



OPEN

# INGRIDKG: A FAIR Knowledge Graph of Graffiti

DATA DESCRIPTOR

Mohamed Ahmed Sherif<sup>1</sup>, Ana Alexandra Morim da Silva<sup>1</sup>, Svetlana Pestryakova<sup>1</sup>,  
Abdullah Fathi Ahmed<sup>1</sup>, Sven Niemann<sup>2</sup> & Axel-Cyrille Ngonga Ngomo<sup>1</sup>

Graffiti is an urban phenomenon that is increasingly attracting the interest of the sciences. To the best of our knowledge, no suitable data corpora are available for systematic research until now. The Information System Graffiti in Germany project (INGRID) closes this gap by dealing with graffiti image collections that have been made available to the project for public use. Within INGRID, the graffiti images are collected, digitized and annotated. With this work, we aim to support the rapid access to a comprehensive data source on INGRID targeted especially by researchers. In particular, we present INGRIDKG, an RDF knowledge graph of annotated graffiti, abides by the Linked Data and FAIR principles. We weekly update INGRIDKG by augmenting the new annotated graffiti to our knowledge graph. Our generation pipeline applies RDF data conversion, link discovery and data fusion approaches to the original data. The current version of INGRIDKG contains 460,640,154 triples and is linked to 3 other knowledge graphs by over 200,000 links. In our use case studies, we demonstrate the usefulness of our knowledge graph for different applications.

## Background & Summary

Graffiti is increasingly attracting the interest of different disciplines like linguistics, art history, anthropology and sociology. Until now, suitable data corpora for systematic research have been lacking. Depending on the research interests, further information about the place where a graffiti is painted, time when it is created, artists who created it, content and language are relevant in addition to the photographs themselves. Although countless photographs of graffiti can be found on the internet, they are not useful for research because the image rights often cannot be determined. Furthermore, the images are usually missing the qualitative metadata containing the aforementioned information about the graffiti.

The research project INGRID (<https://www.uni-paderborn.de/forschungsprojekte/ingrid/projekt>) “Information System Graffiti in Germany” closes this gap. In the summer of 2012, the idea was born to establish an interdisciplinary graffiti database within the framework of an interdisciplinary cooperation between the art historians at the *Karlsruhe Institute of Technology* (KIT) and the *Paderborn University*. Since the practice of graffiti writing is often characterized by a mix of pictorial and written language, an interdisciplinary orientation (i.e., linguistics and art history) of the project was obvious.

Since 2016, INGRID has been funded by the German research foundation DFG (<https://www.dfg.de>). The first project phase, from 01.04.2016 to 30.06.2019, aimed at building a terminology to systematically document and analyze graffiti in Germany. The categories within the graffiti terminologies are created based on our analysis of the available graffiti. The creation of our graffiti terminology has been an ongoing process which is reviewed constantly. Within the first phase of the project, over 40,000 graffiti from the *police department of Mannheim* and the private collection of *Peter Kreuzer* from *Munich* were annotated and made accessible for research purposes.

In the current second project phase, from 01.07.2020 to 30.06.2023 (also funded by DFG), photos provided by the *police department in Munich* as well as photos from the private collector *Dirk Kreckel* among others have been annotated. Dirk Kreckel has been photographing graffiti at important hot spots in Germany like Hamburg, Wiesbaden and Dortmund for decades. The aim of the second phase of INGRID is to test and use crowd-sourcing methods for collecting and annotating the graffiti images. Therefore, from 01.05.2022 to 30.06.2022, the citizens of the city of Paderborn had the opportunity to document photos of graffiti in the city area via the *LingScape* application (<https://lingscape-app.uni.lu>). The photos then were transferred to our database of INGRID.

<sup>1</sup>DICE Research Group, Department of Computer Science, Paderborn University, Paderborn, 33098, Germany.

<sup>2</sup>Institute for German Language and Comparative Literature, Paderborn University, Paderborn, 33098, Germany.

✉e-mail: [mohamed.sherif@upb.de](mailto:mohamed.sherif@upb.de); [afaahmed@mail.upb.de](mailto:afaahmed@mail.upb.de)

The INGRID database currently consists of approximately 136,000 photos from Mannheim, Munich, Cologne, Berlin, Leipzig, Hamm, Dortmund, Münster and Paderborn from the years 1983 to 2019. The images of the graffiti are collected from archives, private property and police inventory. A continuous expansion of the graffiti sources is planned within the project. By building this extensive graffiti database, INGRID is preserving ephemeral art by documenting, digitizing and analyzing graffiti in the public space. Based on the extensive, secure and high-quality research data within the INGRID, the developments and changes in the phenomenon of graffiti can be explored over longer periods of time. INGRID enable the researchers to investigate the visual aesthetics, specific scripturality, notational iconicity, urban location, social function and meaning of each graffiti. After the annotation of the content of the graffiti images, the resulted data is transferred to a permanent domain that provide interfaces and connections to existing standards such as the *Integrated Authority File* (GND). Through connections to the major meta-databases and data infrastructure programs of the participating subjects, such as Prometheus (<https://prometheus.io>) and CLARIN-D (<https://www.clarin-d.net/de/>), an interdisciplinary networking is achieved. As of September 2022, 87,909 annotated graffiti are currently accessible for the public use from the project web site. For accessing our graffiti database, users need to log in to (<https://media.uni-paderborn.de/>). The original images of the graffiti may be used publicly under the license Attribution-ShareAlike 4.0 International (CC BY-SA 4.0), (<https://creativecommons.org/licenses/by-sa/4.0/>). On the other hand, our generated metadata are available under the CC0 license (<https://creativecommons.org/share-your-work/public-domain/cc0/>). Our knowledge graph, INGRIDKG, is publicly available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. See (<https://creativecommons.org/licenses/by/4.0/>) for more details. By default, the provided images contain a watermark, but we can provide the images in higher resolution and without a watermark on demand in response for individual requests.

In this paper, we present INGRIDKG, a comprehensive RDF knowledge graph of annotated graffiti images. Our knowledge graph follows the Linked Data lifecycle<sup>1</sup>. We provide detailed representation of the annotated graffiti in RDF including properties like graffiti's text, location, creators and annotators. Resources such as graffiti and sprayer crews augment the original data and make it easier to process for the sake of question answering and machine learning.

Our knowledge graph abides by the *FAIR principles*<sup>2</sup>. It is **Findable** by virtue of being annotated with rich metadata and indexable by search engines. We make it **Accessible** by providing our data via an RDF dump download (<https://doi.org/10.5281/zenodo.7560242>) as well as (<https://hobbitdata.informatik.uni-leipzig.de/INGRID/>). All resources within the knowledge graph are dereferenceable via HTTP IRIs, which can be accessed via LodView (<https://lodview.it/>) or via the INGRIDKG's SPARQL endpoint (<https://graffiti.data.dice-research.org/sparql>). For instance, Fig. 1 shows the LodView's visualization of the resource (*grfr*:64681). We also make our knowledge graph **Interoperable** by employing standard vocabularies, e.g., for crews, crew members and annotators, as well as through the links to 3 knowledge graphs including DBPEDIA, WIKIDATA and LINKEDGEODATA. Finally, we make INGRIDKG **Reusable** by associating the data with clear provenance and licensing information as well as by reusing popular vocabularies such as *schema* and *rdfs* ourselves.


INGRIDKG allows to obtain a deeper analysis of textual data that cannot be carried out with the user search in our database. For example, a SPARQL query can show which letters are most often used in sprayer names. The answer to this question is important because the letters in a sprayer or crew name in a graffiti writing are based on principles of visual design. A graffiti writing can be described as a stylized signature on the wall that is based on letters, local environment and stylistic elements. We could measure the degree of complexity of a graffiti by computing the number of letters, the use of different colors, or the number of style elements. These correlations between subcultural community, public space, time, quantity and quality are important questions for empirical studies in different disciplines of science such as art history, onomastics, linguistics, or social science. Other potential *use cases* of our knowledge graph include:

- Finding all graffiti painted by a certain crew and showing relationships of a social practice
- Analyzing local subcultural practices by researching the style in a certain city
- Investigating language economy and the complexity of written language
- Focusing on language and environment by linking information about the city

There are a number of other projects focused on the graffiti data collection and annotation. For instance, the project SPRAYCITY (<https://spraycity.at/>) archives a digital collection regarding graffiti in Austria. The INDIGO project (<https://projectindigo.eu/>) aims to provide the groundwork for thoroughly documenting, disseminating and analyzing the over 13 km of unbroken graffiti along Vienna's *Donaukanal* (English: Danube Canal) in the next ten years. Finally, the STORM project (<https://www.heritageresearch-hub.eu/project/storm/>) aims at safeguarding of cultural heritage through technical and organisational resource management.

## Methods

**Graffiti annotation.** The INGRID database is hosted by the EasyDB. The database contains over 130,000 photographs of graffiti. The images were provided to INGRID by different sources. A large part of the pictures comes from the police departments in Mannheim, Cologne and Munich. Another part consists of photographs from private collections and public archives. In order to be able to use the images, a contract is made between the Paderborn University and the copyright holders, which transfers the rights for public use of the images to the INGRID project. This agreement allows INGRID users to publish images as well as images' metadata publicly (licensing: CC BY-SA 4.0) for the original images, CC0 for our generated metadata and (CC BY 4.0) for INGRIDKG. In addition to the collection of images, a central task of the INGRID project is to develop standards for research and analysis of graffiti. The development of these standards led to an ontology that captures constitutive aspects of graffiti (i.e., imagery, scripturality, context, locality, materiality, time, actors, etc.). The ontology also takes into



grfo:creationDate	2017-08-19T13:42:16Z	xsd:dateTime
grfo:hasCarrierMedium	Brücke	
grfo:hasCharacterType	Buchstaben	
grfo:hasColour	Blauton Grünton Rotton	
grfo:hasEmbeddedGraffiti	Signatur/en	
grfo:hasFigureStyle	Linear	
grfo:hasImageLayer	Graffiti	
grfo:hasItem	Worm @en	
grfo:hasLanguage	en - Englisch	
grfo:hasLetterStyle	Simple Style	
grfo:hasMotive	Tier	
grfo:hasOriginalPhoto	1	xsd:boolean
grfo:hasSurface	Beton	

**Fig. 1** Screenshot of the resource page `grfr:64681`.

account categories that are relevant to the graffiti scene itself. These include the distinction of certain letter styles (e.g., simple style, wild style and bubble style), the typology (e.g., pieces, tags and characters), and the use of typical stylistic elements (e.g., crowns, arrows and outline). This strategy led to an ontology based on folk categories as well as linguistic categories or art historical terms. The information about the location, the recording date and the origin of the images is automatically imported by the *Information and Media Technologies Centre (IMT)* (<https://imt.uni-paderborn.de/en/>) of Paderborn University during the import process from the original mediums such as DVD, CD and memory sticks. When imported into the database, each image is assigned a `systemID`, which allows users to uniquely cite images. The `systemID` allows other system users to track and repeat the analyses. The annotation is done by six annotators in two different teams: one in Paderborn (linguistics) and the other in Karlsruhe (art history). The team in Paderborn focuses on aspects of scripturality such as the analysis of language, grammar and semiotics. The team in Karlsruhe focuses on aspects of imagery such as the shape of letters, the design of figurative elements, and the choice of colors.

**The annotation process details.** The annotation process is carried out by selecting the proper categories and subcategories for each graffiti by the annotators. For example, the category *type* contains a number of different subcategories (e.g., *tag*, *stencil*, *piece*). Table 1 shows a sample of the annotations' categories and its respective

Categories	Subcategories
Technology	Pencil, Roller, Pen, Pressure scratch, chalk, Stencil, Spray can, Other
Carrier medium	Trailer, Vending machine, Railway line, Tree/Plant, Ground, Mailbox, Bridge, Bus, Container, Downpipe, Window/Blind, Railing/Handrail, Closed interior, Freight train, Hall of Fame, Stop noise barrier, Truck, Mast, Waste/Disposal container, Park bench, Passenger train, Car, Column/Pillar/Support, Display, Case Sign, Play/Sports equipment, Sticker, Road salt container, Electricity/Distribution box, Telephone box, Gate, Stairs, Door, Subway/S-Bahn Subway/S-Bahn/Passenger train - inside underpass, Wall Fence, Other, Undefined
Surface	Asphalt, Concrete, Glass, Wood, Ceramics, Plastic, Metal, Natural stone, Paper/Cardboard, Plaster, Exposed, Aggregate, Concrete Brick, Other, Undefined
Type	Character, Comment, Co-Construction, Composition, Piece/Writing/Style, Sketch, Saying/Concept call, Stencil, Tag(s), Throw Up, Otherwise
Subject area	Drugs, Erotic, Film, Peace, Football, Sports, Violence, Culture Art, Love, Music, Politics, Police, Racism, Religion, School, Death, Environment, Behavioural standard, Economy, Other, None

**Table 1.** Annotation categories.

File name	Description
public_rdfGraffiti.ttl	Contains the public version of INGRIDKG, including the graffiti resources and the linking information.
ontology_v3.ttl	Contains the ontology of INGRID's KG.
void.ttl	Contains metadata on the dataset
ingrid_similar_images_public_1.ttl	First part of the similarity scores file.
ingrid_similar_images_public_2.ttl	Second part of the similarity scores file.
ingrid_similar_images_public_3.ttl	Third part of the similarity scores file.
ingrid_public_images.tar.gz	Compressed file with all the public images of the graffiti, where the name of each image file reflects the graffiti ID in the KG.

**Table 2.** List of files at our ZENODO repository.

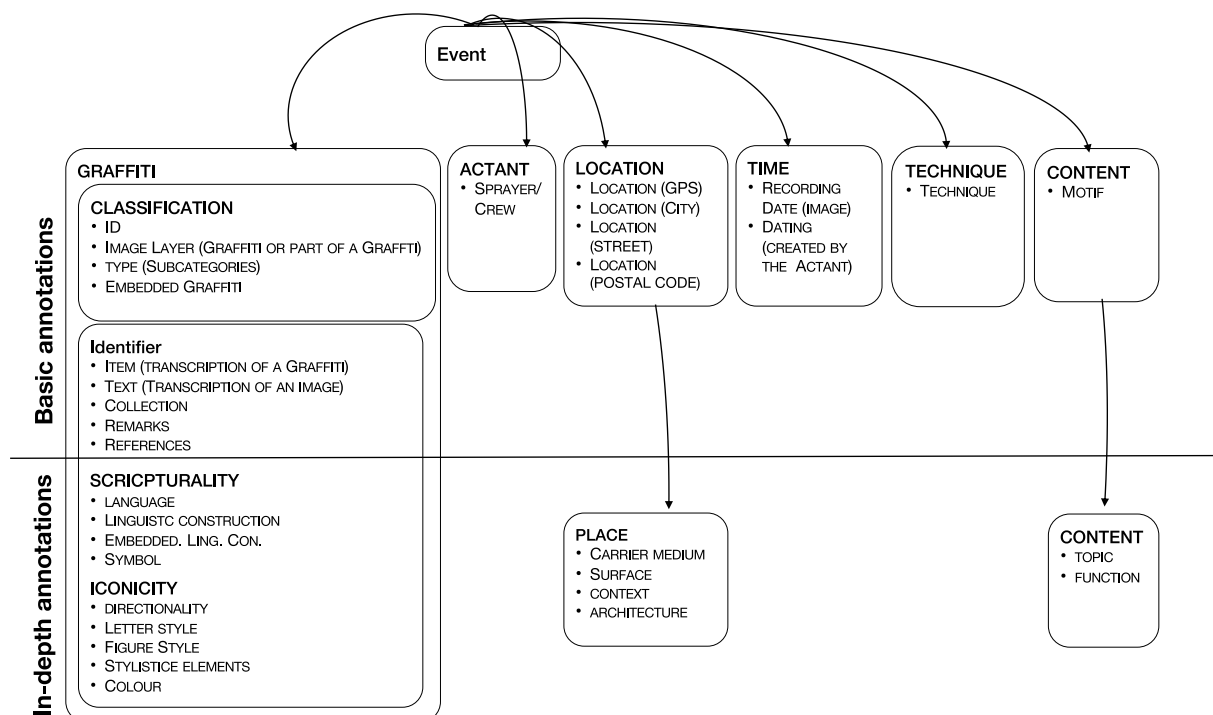
Name	INGRIDKG
Zenodo archive	<a href="https://doi.org/10.5281/zenodo.7560242">https://doi.org/10.5281/zenodo.7560242</a>
INGRIDKG license	Creative Commons Attribution 4.0 International ( <a href="https://creativecommons.org/licenses/by/4.0/">https://creativecommons.org/licenses/by/4.0/</a> )
Example resource	<a href="https://graffiti.data.dice-research.org/resource/64681">https://graffiti.data.dice-research.org/resource/64681</a>
INGRIDKG dump	<a href="https://hobbitdata.informatik.uni-leipzig.de/INGRID/">https://hobbitdata.informatik.uni-leipzig.de/INGRID/</a>
Archived dump	<a href="https://hobbitdata.informatik.uni-leipzig.de/INGRID/archive">https://hobbitdata.informatik.uni-leipzig.de/INGRID/archive</a>
Sparql endpoint	<a href="https://graffiti.data.dice-research.org/sparql/">https://graffiti.data.dice-research.org/sparql/</a>
INGRIDKG graph	<a href="https://hobbitdata.informatik.uni-leipzig.de/INGRID/public_rdfGraffiti.ttl">https://hobbitdata.informatik.uni-leipzig.de/INGRID/public_rdfGraffiti.ttl</a>
Void file	<a href="https://hobbitdata.informatik.uni-leipzig.de/INGRID/void.ttl">https://hobbitdata.informatik.uni-leipzig.de/INGRID/void.ttl</a>
Version date	July 5, 2022
Version Number	4.0
Ontology	<a href="https://hobbitdata.informatik.uni-leipzig.de/INGRID/ontology_v3.ttl">https://hobbitdata.informatik.uni-leipzig.de/INGRID/ontology_v3.ttl</a>
	<a href="https://doi.org/10.5281/zenodo.7560242">https://doi.org/10.5281/zenodo.7560242</a>
	<a href="https://github.com/dice-group/ingrid.github.io/blob/main/ontology_v3.ttl">https://github.com/dice-group/ingrid.github.io/blob/main/ontology_v3.ttl</a>
Source code	<a href="https://github.com/dice-group/Ingrid">https://github.com/dice-group/Ingrid</a>
Software license	GPL 3.0 ( <a href="https://www.gnu.org/licenses/gpl-3.0">https://www.gnu.org/licenses/gpl-3.0</a> )
Graffiti images archive	<a href="https://doi.org/10.5281/zenodo.7759189">https://doi.org/10.5281/zenodo.7759189</a>
	<a href="https://hobbitdata.informatik.uni-leipzig.de/INGRID/ingrid_public_images.tar.gz">https://hobbitdata.informatik.uni-leipzig.de/INGRID/ingrid_public_images.tar.gz</a>
Graffiti images license	CC BY-SA 4.0, ( <a href="https://creativecommons.org/licenses/by-sa/4.0/">https://creativecommons.org/licenses/by-sa/4.0/</a> )
Database	<a href="https://media.uni-paderborn.de/">https://media.uni-paderborn.de/</a>
Annotation manuals	<a href="https://dice-group.github.io/ingrid.github.io/">https://dice-group.github.io/ingrid.github.io/</a>
Annotations license	CC0, ( <a href="https://creativecommons.org/share-your-work/public-domain/cc0/">https://creativecommons.org/share-your-work/public-domain/cc0/</a> )

**Table 3.** Technical details of INGRIDKG.

sub-categories. Please refer to our online annotation manual (<https://dice-group.github.io/ingrid.github.io/>) for the complete categories and subcategories table. An overview of the central categories are given in 2 In addition to the selection of fixed categories, our annotators are able also to fill out open text fields when annotating a graffiti (e.g., *item*, *text*, *remarks*) The category *text* is systematic representation of all graffiti in written form, which is very usable as graffiti is often difficult to read due to its deformation of letters. The open text field *remarks* gives useful information with some background knowledge about the topic or the motif that can help the users to understand the meaning of a graffiti. In order to keep the Inter-Annotator Agreement (IAA) high, the annotation

	29.06.2021	30.08.2021	28.02.2022	05.07.2022	Latest
Distinct number of all resources	640,536	803,675	835,597	841,167	840,420
Distinct number of graffiti	100,758	125,962	130,692	130,831	130,689
Distinct number of sprayer crews	15,044	16,656	21,616	23,380	23,272
Distinct number of graffiti symbols	0	0	0	46	46
Distinct number of crews	348	348	348	0	0
Distinct number of crew members	581	581	581	948	948
Distinct number of cities	149	160	165	178	178
Distinct number of image files	523,242	659,470	681,378	683,957	682,748
Distinct number of collections	0	0	0	451	451

**Table 4.** INGRIDKG statistics.



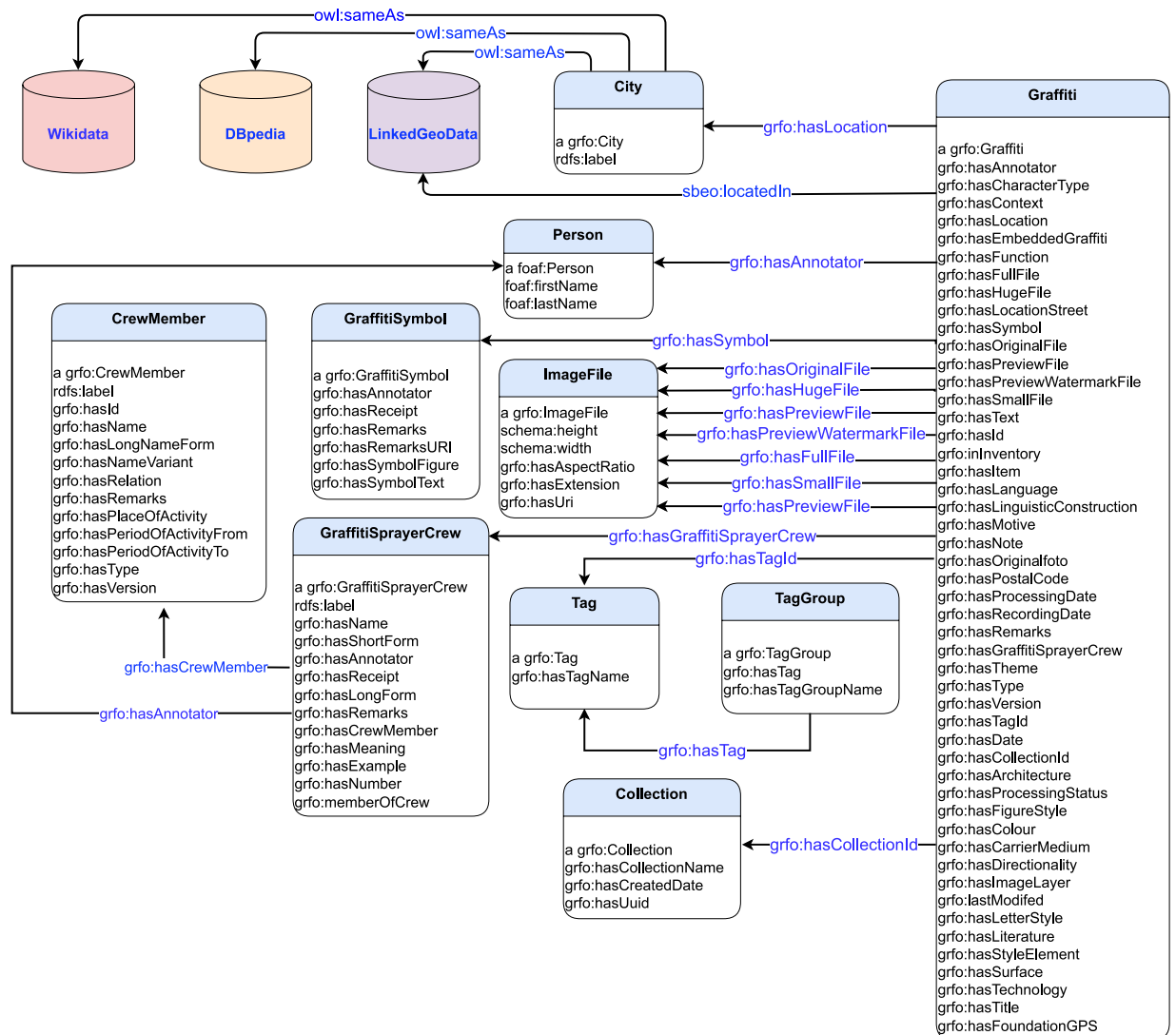
**Fig. 2** Annotations categories.

```

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix lgdt: <http://linkedgedata.org/triplify/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix sbeo: <https://qasimkhalid.github.io/SBEO/#> .
@prefix grfr: <https://graffiti.data.dice-research.org/resource/> .
@prefix grfo: <https://graffiti.data.dice-research.org/ontology/> .
  
```

**Fig. 3** List of all used vocabularies in INGRIDKG.

on each individual category and each open text field is based on specific rules defined in an annotation manual. The annotation manual contains all names of the used categories and terms with the corresponding classification criteria. For example, the category *piece* is defined as a large, multicolored graffiti with areal letters. The central element of a piece is usually decorative with typical stylistic elements and can include figurative representations (characters). Smaller forms of graffiti are often embedded in a piece (e.g., tags, comments or dating). For more information a German and an English version of the complete manual is available on the project homepage and can be downloaded via the following links: German manual ([https://www.uni-paderborn.de/fileadmin/ingrid/INGRID\\_Manual\\_Oktober\\_2019.pdf](https://www.uni-paderborn.de/fileadmin/ingrid/INGRID_Manual_Oktober_2019.pdf)), English manual ([https://github.com/dice-group/ingrid.github.io/blob/main/INGRID\\_Manual\\_Oktober\\_2019en.pdf](https://github.com/dice-group/ingrid.github.io/blob/main/INGRID_Manual_Oktober_2019en.pdf)). Even though the manual defines clear rules for annotation, in some cases graffiti appear that cannot always be clearly assigned to a category. In such cases, the annotation of a



**Fig. 4** UML class diagram for the INGRIDKG ontology.

graffiti is clarified within our weekly meeting. Once the annotation of the images of the graffiti is completed, the images are made available for search in our instance of the EASYDB database (<https://media.uni-paderborn.de/>). Moreover, we provide a public full images dump via the HOBBIT data server at [https://hobbitdata.informatik.uni-leipzig.de/INGRID/ingrid\\_public\\_images.tar.gz](https://hobbitdata.informatik.uni-leipzig.de/INGRID/ingrid_public_images.tar.gz).

**Knowledge graph generation.** We start the process of our knowledge graph generation by creating our INGRID ontology. In particular, we define the basic classes (i.e., graffiti, annotator, crew and crew members) and predicates associated with each of the class's entities. We detail our ontology generation process in the next section. Based on our ontology, we developed *Python* scripts for converting the database records of the graffiti images into our RDF INGRIDKG. We then link our created INGRIDKG with external knowledge graphs. Currently, we link INGRIDKG to the knowledge graphs of LINKEDGEODATA, WIKIDATA and DBPEDIA. We details our linking process in the *Linking Section*.

### Data Records

Our knowledge graph, INGRIDKG, is available at the open science portal ZENODO<sup>3</sup>. Currently, ZENODO's repository mirrors the data present in our SPARQL endpoint (<https://graffiti.data.dice-research.org/sparql/>) and in our data dump (<https://hobbitdata.informatik.uni-leipzig.de/INGRID>). In particular, we include the public version of INGRIDKG with graffiti resources and respective linking information, the ontology file, the dataset metadata expressed with the standardized VOID vocabulary, the similarity scores between the different graffiti resources in the form of reified statements, and the images of the publicly available graffiti. Furthermore, we provide a README file with a short description of each file within the repository. In Table 2, we list the files available at our ZENODO<sup>3</sup> together with its respective descriptions. In this section, we begin by describing the **structure** of our ontology, where we introduce each of its classes, name spaces and data model. We then introduce our knowledge graph linking procedures. Finally, we detail how we automate our knowledge graph generation process.

```

grfr:64681 a      grfo:Graffiti ;
  grfo:creationDate      "2017-08-19T13:42:16+00:00"^^xsd:dateTime ;
  grfo:hasCarrierMedium  "Brücke" ;
  grfo:hasCharacterType  "Buchstaben" ;
  grfo:hasCollectionId   grfr:collection_444 , grfr:collection_756 , grfo:collection_655 ;
  grfo:hasColour         "Grünton" , "Blauton" , "Rotton" ;
  grfo:hasEmbeddedGraffiti "Signatur/en" ;
  grfo:hasFigureStyle   "Linear" ;
  grfo:hasFullFile      grfr:Full_63233 ;
  grfo:hasGraffitiSprayerCrew grfr:RICO ;
  grfo:hasHugeFile      grfr:Huge_63233 ;
  grfo:hasImageLayer    "Graffiti" ;
  grfo:hasItem          "Worm"@en ;
  grfo:hasLanguage      "en - Englisch" ;
  grfo:hasLetterStyle   "Simple Style" ;
  grfo:hasLocation      grfr:München ;
  grfo:hasMotive        "Tier" ;
  grfo:hasPreviewFile   grfr:Preview_63233 ;
  grfo:hasPreviewWatermarkFile grfr:PreviewWatermark_63233 ;
  grfo:hasSmallFile     grfr:Small_63233 ;
  grfo:hasSurface       "Beton" ;
  grfo:hasTagId         grfr:tag_11 , grfr:tag_8 , grfr:tag_20 , grfr:tag_18 ;
  grfo:hasTechnology    "Sprühdose" ;
  grfo:hasText          "Worm | RiCO"@en ;
  grfo:hasTheme         "sonstiger" ;
  grfo:hasTitle         "FS-NL-KREU-01-06-0005" ;
  grfo:hasType          "Komposition" ;
  grfo:hasVersion       "9"^^xsd:nonNegativeInteger ;
  grfo:inInventory      "Stadtarchiv München, Sammlung Kreuzer" ;
  grfo:lastModified     "2022-06-14T15:32:32+00:00"^^xsd:dateTime .

```

**Fig. 5** Example graffiti resource.

```

grfr:110 a      grfo:GraffitiSprayerCrew ;
  grfo:hasReceipt      "IMG_8312"^^xsd:string ;
  grfo:hasCrewMember  grfr:SHOP ;
  rdfs:label           "110"^^xsd:string .
  grfo:hasLongForm     "110 CREW"^^xsd:string ;
  grfo:hasName         "110"^^xsd:string .

```

**Fig. 6** Example of a crew representation.

```

grfr:TUMOR a      grfo:CrewMember ;
  rdfs:label       "TUMOR"^^xsd:string;
  grfo:hasName     "TUMOR"^^xsd:string ;
  grfo:hasLongForm "Mister Tumor one"^^xsd:string ;
  grfo:hasReceipt  "IMG_11088 " ;
  grfo:hasRemarks "z.T. Kringel an R, der wie ein e aussieht; Sprachen: de, en, es"^^xsd:string .

```

**Fig. 7** Example of a crew member.

```

grfr:nearDuplicate_51855_64681 a      rdf:Statement ;
  grfo:hasDHashScore 5;
  rdf:subject         grfr:51855 ;
  rdf:predicate       grfo:nearDuplicate ;
  rdf:object          grfr:64681 .

```

**Fig. 8** Example of near duplicate resources of grfr:64681.

Our knowledge graph creation process is implemented in *Python 3.6* with *RDFLib 5.0.0* (<https://github.com/RDFLib/rdfliib>). We make our source code publicly available (<https://github.com/dice-group/Ingrid>) to ensure the *reproducibility* of our results and the rapid conversion of novel graffiti database versions. We present some statistics regarding the increasing size of the INGRIDKG's resources over time in Table 4.

**Ontology design.** The ontology behind our knowledge graph is derived from the source from which it was extracted, i.e., the database of the graffiti image annotations. The ontology is designed to enable search, question answering and machine learning.

As part of our continuous **ontology refinement** work, we fused the instances of the `Crew` class into the `SprayerCrew` class, which is the reason of the disappearance of the `Crew` class instances and the increased number of the instances of the `SprayerCrew` class starting from the 05.07.22 version. Moreover, we added the `Collection` class in the same version Fig. 2.

**RDF namespaces.** To facilitate the reusability of our knowledge graph, we represent our data in widely used vocabularies and namespaces as shown in Fig. 3.

**RDF data model.** Figure 4 shows important classes (e.g., graffiti, crew, crew member, person, image file and city) as well as predicates (e.g., graffiti's location, annotator and text within a graffiti).

**Graffiti.** We represent graffiti as instance of class `grfo:Graffiti`. Each graffiti instance contains information regarding the graffiti's provenance, contents, creators and annotators. Important attributes regarding graffiti contents include text, character type, language, color, theme, text direction, carrier medium and context. For each graffiti, we store **provenance information**. In particular, using the predicate `grfo:inInventory`, we explicitly state the original inventory where the raw image file of the graffiti come from. INGRIDKG also allows referencing to the original image raw files as well as the place and time where/when the image is taken. Moreover, we store the time when we annotate the resource. The URIs of our generated `Graffiti` resources follow the format <https://graffiti.data.dice-research.org/resource/graffitiId> where `graffitiId` is the unique id for each graffiti within INGRIDKG. For example, the original image of the graffiti resource presented in Fig. 5 is created on 2017-08-19 at 13:42:16 o'clock from the inventory of `Stadtarchiv Munchen, Sammlung Kreuzer` and the last annotation work of the image is on 2022-06 at 14:00 to 15:32:327 o'clock.

**Collection.** Each graffiti in INGRIDKG is linked to the collection it came from. There are five different types of collections: (i) Police Department Mannheim, (ii) Police Department Munich, (iii) Police Department Cologne, (iv) The collection of Dirk Kreckel and (v) The collection of Peter Kreuzer.

**Annotator.** Our annotators are represented in the FOAF (<http://xmlns.com/foaf/spec/http://xmlns.com/foaf/spec/>) vocabulary. In particular, we include each annotator's first, middle and last name as well as mail address and institute. Each graffiti instance is linked to its respective annotators via the `grfo:hasAnnotator` predicate (as shown in Fig. 4). For data protection reasons, we can not provide examples of the annotator resources. (Note that, that is the only part of our knowledge graph that is not publicly available due to data protection reasons.)

**Crew.** Each graffiti is linked to the crews that painted it via the `grfo:hasGraffitiSprayerCrew` predicates. Each crew instance include information related to crew name in short and long form, members and any spacial notes. Each crew is also linked to its annotator via the `grfo:hasAnnotator` predicate. In Fig. 6, we provide an example of a crew resource.

**Crew members.** Each crew contains one or more crew members. Therefore, we store crew members' corresponding information as instance of class `grfo:CrewMember`. In particular, we store for each crew member his/her name, name variants, abbreviations and remarks. See Fig. 7 for an example of a crew member resource. **Image files.** For each graffiti, we store its original image files in various qualities and sizes. Each `grfo:Image-File` instance include the image file's height, width, aspect ratio, extension and URI.

**Linking.** We link our dataset to other data sources to ensure its reusability and integrability as well as to improve its use for search, question answering and structured machine learning. We generate links from our graffiti resources to publicly available related knowledge bases. In our linking process, we rely on Limes<sup>4</sup> as it is a state-of-the-art declarative link discovery framework with open source implementation that can be easily adopted and configured. We manually created a Limes configuration file for each linking task. All Limes configuration tasks are available from the project web site (<https://github.com/dice-group/Ingrid>).

**Linking graffiti.** We link INGRIDKG's resources of type `grfo:Graffiti` to LINKEDGEODATA resources of type `lgdo:Building`. In particular, we configure Limes to declare a link (i.e., `sbeo:locatedIn`) if both city name and street name of the `grfo:Graffiti` and the `lgdo:Building` are matched using the *jaccard* similarity of at least 90%. Following the same method, we also link our resources of type `grfo:Graffiti` to LINKEDGEODATA resources of type `lgdo:HighwayThing`.

**Linking cities.** Each graffiti resource in our dataset (i.e., of type `grfo:Graffiti`) is linked to a city resource (i.e., of type `grfo:City`) using the predicate `sbeo:locatedIn` as shown in Fig. 4. We link `grfo:City` resources from INGRIDKG to the LINKEDGEODATA knowledge graph using the `owl:sameAs` predicates. In particular, we link resources of type `grfo:City` from our knowledge graph to resources of type `lgdo:City` from LINKEDGEODATA dataset. In particular, we configure Limes to declare a match once the *levenshtein* similarity among the `rdfs:label` of a `grfo:City` resource from INGRIDKG and a `lgdo:City` resource from LINKEDGEODATA is above 90%. In a similar way, we link the INGRIDKG's resources of a `grfo:City` to DBPEDIA resources of type `dbo:Location`. We configured Limes to declare a link once the *jaccard* similarity between the `rdfs:label` of `grfo:City` resource and a `dbo:Location` resource is above 60%. Finally,



Links	Annotators			Mutual agreement
	I	II	III	
grfr:Leverkusen owl:sameAs wd:Q2938.	✓	✓	✓	✓
grfr:Hirschberg owl:sameAs wd:Q32058833.	✗	✓	✗	✗
grfr:Hirschberg owl:sameAs wd:Q630383.	✓	✓	✓	✓
grfr:Hirschberg owl:sameAs wd:Q468337.	✓	✓	✗	✓
grfr:Weinheim owl:sameAs wd:Q7050.	✓	✓	✓	✓
grfr:Bielefeld owl:sameAs wd:Q2112.	✓	✓	✓	✓
grfr:Osnabrück owl:sameAs wd:Q2916.	✓	✓	✓	✓
grfr:Osnabrück owl:sameAs wd:Q33158934.	✓	✓	✓	✓
grfr:Darmstadt owl:sameAs wd:Q2973.	✓	✓	✓	✓
grfr:Regensburg owl:sameAs wd:Q2978.	✓	✓	✓	✓
grfr:Regensburg owl:sameAs wd:Q2978.	✓	✓	✓	✓
grfr:Unna owl:sameAs wd:Q3949.	✓	✓	✓	✓
grfr:Olching owl:sameAs wd:Q32206345.	✓	✓	✓	✓
grfr:Olching owl:sameAs wd:Q178486.	✓	✓	✓	✓
grfr:Essen owl:sameAs wd:Q2066.	✓	✓	✓	✓
grfr:Hockenheim owl:sameAs wd:Q32059421.	✓	✓	✗	✓
grfr:Dresden owl:sameAs wd:Q1731.	✓	✓	✓	✓
grfr:Oberhausen owl:sameAs wd:Q32200661.	✓	✓	✓	✓
grfr:Oberhausen owl:sameAs wd:Q32200649.	✓	✓	✓	✓
grfr:Neuss owl:sameAs wd:Q2948.	✓	✓	✓	✓
grfr:Leutershausen owl:sameAs wd:Q389945.	✓	✓	✓	✓
grfr:Heidelberg owl:sameAs wd:Q2966.	✓	✓	✓	✓
grfr:Leverkusen owl:sameAs wd:Q2938.	✓	✓	✓	✓
grfr:Schwerte owl:sameAs wd:Q6863.	✓	✓	✓	✓
grfr:Wolfratshausen owl:sameAs wd:Q503160.	✓	✓	✓	✓
grfr:Hemsbach owl:sameAs wd:Q81012.	✓	✓	✓	✓
grfr:Werdohl owl:sameAs wd:Q5575.	✓	✓	✓	✓
grfr:Ketsch owl:sameAs wd:Q32064800.	✓	✓	✓	✓
grfr:Dossenheim owl:sameAs wd:Q31971934.	✓	✓	✓	✓
grfr:Bremen owl:sameAs wd:Q1209.	✓	✓	✗	✓
grfr:Bremen owl:sameAs wd:Q24879.	✓	✓	✓	✓
grfr:Paderborn owl:sameAs wd:Q2971.	✓	✓	✓	✓
grfr:Bobenheim-Roxheim owl:sameAs wd:Q31916736.	✓	✓	✓	✓
grfr:Mannheim owl:sameAs wd:Q2119.	✓	✓	✓	✓

**Table 5.** Manual annotation results of links generated by Limes among cities of INGRIDKG and WIKIDATA.

we linked our grfo:City resources to the ones from WIKIDATA of type wd:Q183. For instance, LIMES discovered 208,942 sbeo:locatedIn links between INGRIDKG and LINKEDGEO DATA, and 65 owl:sameAs links between INGRIDKG and WIKIDATA.

**Linking postal codes.** We link each of the graffiti with known postal code with LINKEDGEO DATA locations with the exact match lgdo:postalCode via the predicate sbeo:locatedIn. As the graffiti are located in Germany and postal codes are not unique worldwide, we configured Limes to restrict the postal codes to the ones in Germany (i.e., postal codes which are managed by the *Deutsche Post*).

**Linking exact duplicated and near duplicated Images.** We linked each graffiti resource in our Knowledge graph to all its image's duplicates and near duplicates. Near duplicate images of graffiti generally exists as a result of one graffiti being captured from different view points, distance, illumination conditions of different resolution. In particular, we used the predicate grfo:nearDuplicate to link each graffiti resource to its exact duplicated and near duplicated resources. For example, we can see Fig. 8 that the graffiti resource grfr:64681 has the near duplicates resources grfr:51855. Technically, we first used the PYTHON package IMAGEDEDUP<sup>5</sup> to encode all the images. We then used image duplication technique DHash (<https://github.com/idealo/imagededup>) to carry out our near duplication experiments.

### Technical Validation

**Annotation validation.** The annotators use a manual in which the rules for the annotation are defined. Cases of uncertainty are discussed in a weekly team meeting in ZOOM. In addition, the annotators use the platform SLACK to exclude any uncertainty. Before the annotation starts, it is checked whether a comparable case has already been annotated in order to keep the inter-annotator agreement as high as possible.

```

SELECT ?text (COUNT(DISTINCT ?graffitiCrew) AS ?cnt) WHERE
{
  ?graffiti grfo:hasGraffitiSprayerCrew ?graffitiCrew .
  ?graffiti grfo:hasItem ?text .
  ?graffitiCrew rdfs:label ?crewName .
}
ORDER BY DESC(?cnt)

```

**Fig. 9** How many crews painted each graffiti?

```

SELECT DISTINCT * WHERE
{
  ?crew a grfo:GraffitiSprayerCrew .
  ?crew grfo:hasCrewMember grfr:REAL .
  grfr:REAL grfo:hasLongForm ?crewMemberLabel .
}

```

**Fig. 10** In which crews does the crew member “REAL” work?

```

SELECT DISTINCT ?text WHERE
{
  ?graffiti grfo:hasItem ?text .
  ?graffiti grfo:hasPostalCode "68159" .
}

```

**Fig. 11** List all graffiti text on the “68159” postal code.

```

SELECT DISTINCT * WHERE
{
  ?graffiti grfo:hasGraffitiSprayerCrew ?crew .
  ?crew a grfo:GraffitiSprayerCrew .
  ?crew rdfs:label "ASKER" .
  ?graffiti grfo:hasItem ?item .
  FILTER( contains(?item, "!") ) .
}

```

**Fig. 12** Find all graffiti drawn by a crew named “ASKER” and item field contains “!”.

```

SELECT DISTINCT (str(?tech) AS ?Technique)
                 (COUNT(?tech) AS ?GraffitiCount)
{
  ?graffiti a grfo:Graffiti .
  ?graffiti grfo:hasTechnology ?tech .
}
ORDER BY DESC (?GraffitiCount)

```

**Fig. 13** Retrieve the count of graffiti per individual technique.

**Ontology validation.** We build our knowledge graph INGRIDKG based on our defined ontology via the usage of our knowledge graph automatic generation scripts (<https://github.com/dice-group/Ingrid>). In particular, we make sure that each of our knowledge graph conversion scripts respect all the defined ontology constraints within our ontology. For instance, all our generation scripts always creates the relation `grfo:creationDate` with the domain of `grfo:Graffiti` and the range of `xsd:dateTime`. Therefore, we do not need to run any ontology validation of our knowledge graph after it is created.

**Linking validation.** As mentioned in the linking section, we use the link discovery framework Limes for generating the links among INGRID and the external knowledge graphs of DBPEDIA, LINKEDGEODATA and WIKIDATA. To validate the automatically generated links by LIMES, we ran a manual annotation of the generated links by at least three annotators. In particular, each annotator manually check the correctness of each link as being True or False. Afterwards, we gather the common voting of the annotator as the mutual agreement

for each link. We then remove all links with a False mutual agreement, if any. Finally, we add the True mutually agreed linked to our knowledge graph. In Table 5, we provides an example of the result of manual annotation of city resources from INGRID and WIKIDATA, where each like is manually checked for correctness by three annotators and the mutual agreement is presented in the last column. Overall, our manual annotators define less than 5% of the links generated by LIMES as wrong and all of them were removed from INGRIDKG.

### Usage Notes

Representing the annotated graffiti images as an RDF knowledge graph promises to facilitate many applications and use cases. In the project web site (<https://dice-group.github.io/ingrid.github.io/>), we provide some commonly asked SPARQL query examples that demonstrate many real world use cases. We outline some of which within this section.

**Data retrieval.** While our database of the INGRID project contains a significant number images, they are not represented in a format optimized for retrieval. By providing INGRIDKG in RDF with a well-defined ontology, we enable the easy retrieval of data with a structured query languages such as SPARQL. For example, in Fig. 9 we show a query to retrieve an ordered list of number of crews participated in painted each graffiti. Another query to retrieve all crews where a specific crew member works is provided in Fig. 10.

**Geographic queries.** Modelling the geographic information associated with each graffiti in our knowledge graph adds a considerable amount of value when to aggregate different pieces of information based on geographic location. For example, Fig. 11 shows a SPARQL query to list all graffiti text for all graffiti within the area of the postal code “68159”.

**Textual data analysis.** Our knowledge graph allows to obtain a deeper analysis of textual data within annotation. For example, Fig. 12 shows a SPARQL query to find all graffiti drawn by a crew named “ASKER” and item field contains “!”.

**Computing statistics.** The reach annotations of the graffiti within INGRIDKG enable the easy computation of statistical aspects of the data. For example, Fig. 13a, we show a SPARQL query that counts the number of graffiti drawn per drawing technique. In Fig. 13b, we show the resulted statistics using the last version of our knowledge graph.

### Data availability

Due to different contractual conditions, only a subset of the photographs of the graffiti are public, with those provided by police departments in Mannheim, Munich and Cologne are not publicly visible. This does not apply to the metadata of the photographs which were compiled by our annotators, all of which are public. All the photographs provided by the private collector *Peter Kreuzer* together with its annotations are publicly available. Note that, some parts of the publicly available graffiti's photographs are distorted for data protection reasons. For data protection reasons, the annotators' names are not publicly available. All the data within INGRIDKG is publicly available under the *Creative Commons Attribution 4.0 International* license.

All our resources are served from one of our servers via **persistent URIs**. For an example resource see (<grfr:64681grfr:64681>). The resource is maintained by the DICE research team (<https://dice-research.org>) as part of the lab's HOBBIT dataset efforts<sup>6</sup>. A 100TB-Server maintained by the Paderborn university's computing center is hosting the knowledge graph. We also provide **dump files** of our dataset for download. INGRIDKG dump files are located on our HOBBIT storage <https://hobbitdata.informatik.uni-leipzig.de/INGRID/> and **archived on Zenodo**<sup>3</sup>. Finally, we publicly serve INGRIDKG via a **SPARQL endpoint** (<https://graffiti.data.dice-research.org/sparql>). Table 3 summarizes all technical details of our knowledge graph pertaining to its availability.

### Code availability

Our source code to generate the new versions of INGRIDKG is publicly available at (<https://github.com/dice-group/Ingrid>) and will be maintained in parallel with the knowledge graph. We provide our source code under the software license of GPL 3.0 (<https://www.gnu.org/licenses/gpl-3.0>).

Received: 24 January 2023; Accepted: 2 May 2023;

Published online: 25 May 2023

### References

1. Ngomo, A.-C. N., Auer, S., Lehmann, J. & Zaveri, A. Introduction to linked data and its lifecycle on the web. In *Reasoning Web International Summer School*, 1–99 (Springer, 2014).
2. Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Scientific data* **3** (2016).
3. Sherif, M. A., da Silva, A. A. M., Pestryakova, S., Ahmed, A. F. & Ngomo, A.-C. N. IngridKG: A FAIR Knowledge Graph of Graffiti. *Zenodo* <https://doi.org/10.5281/zenodo.7759189> (2023).
4. Ngonga Ngomo, A.-C. *et al.* LIMES - A Framework for Link Discovery on the Semantic Web. *KI - Künstliche Intelligenz, German Journal of Artificial Intelligence - Organ des Fachbereichs "Künstliche Intelligenz" der Gesellschaft für Informatik e.V.* (2021).
5. Jain, T., Lennan, C., John, Z. & Tran, D. *Imagededup*. <https://github.com/idealo/imagededup> (2019).
6. Röder, M., Kuchelev, D. & Ngonga Ngomo, A.-C. Hobbitt: A platform for benchmarking big linked data. *Data Science* 1–21 (2019).

### Acknowledgements

This work has been supported by the German Research Foundation (DFG) within the project INGRID under the grant no NG105/7-3. We acknowledge support for the publication costs by the Open Access Publication Fund of Paderborn University.

### Author contributions

*Mohamed Ahmed Sherif* analysed the data and designed the knowledge base ontology and supervised the work. *Svetlana Pestryakova* carried out the first version of the RDF data transformation and linking tasks. *Ana Alexandra Morim da Silva* updated the RDF data transformation pipeline. *Abdullah Fathi Ahmed* carried out the data integration tasks. *Sven Niemann* provided his expertise regarding graffiti annotation and the `EasyDB` instance of INGRID. *Axel-Cyrille Ngonga Ngomo* supervised the work. All authors contributed to the text of the article, read and approved the final manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to M.A.S. or A.F.A.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023