CLEOPATRA

Cross-lingual Event-centric Open Analytics Research Academy 2020

Marie Sklodowska-Curie Innovative Training Networks
(H2020-MSCA-ITN-2018)
Grant Agreement No: 812997

Open Event Knowledge Graph Version 1.0

Autors: Maria Maleshkova (UBO), Elena Demidova (LUH), Simon Gottschalk (LUH), Endri Kacupaj (UBO)

Abstract. This white paper describes the first version of Open Event Knowledge Graph, created by and used in the Cleopatra project. It also includes the contributions of the individual ESR projects towards realisation of this KG.

1. Introduction

Knowledge Graphs (KG) have been increasing in use and importance during the past few years. Knowledge Graphs are a specific type of data storing solutions. Two common definitions used are for instance that a KG is a “a collection of points and lines connecting some (possibly empty) subset of them” (Wolfram MathWorld, 15 Oct 2018) or a KG is “a collection of vertices and edges that join pairs of vertices” (Merriam-Webster, 15 Oct 2018). As it can be seen in Figure 1. In terms of being a current and impactful technology for data storage, KGs are expected to continue to play a major role in the context of data use during the upcoming five to then year.

In the context of the Cleopatra project, knowledge graphs play a major role, since they are the main solution to storing, integrating and processing data. In particular, all ESR projects contribute to the Open Event Knowledge Graph either by contributing extracted data and benchmarks, or by participating in data evaluation and analytics in the context of their studies. In this white paper we discuss the main characteristics of the Open Event Knowledge Graph – multilingual, event-centric, temporal. Furthermore, we describe the architectural solution for the OEKG and what requirements it needs to fulfill. We also list all the datasets that the ESRs are contributing to the KG and go into detail on the first version of the OEKG. We conclude with the planned future work and a short summary.

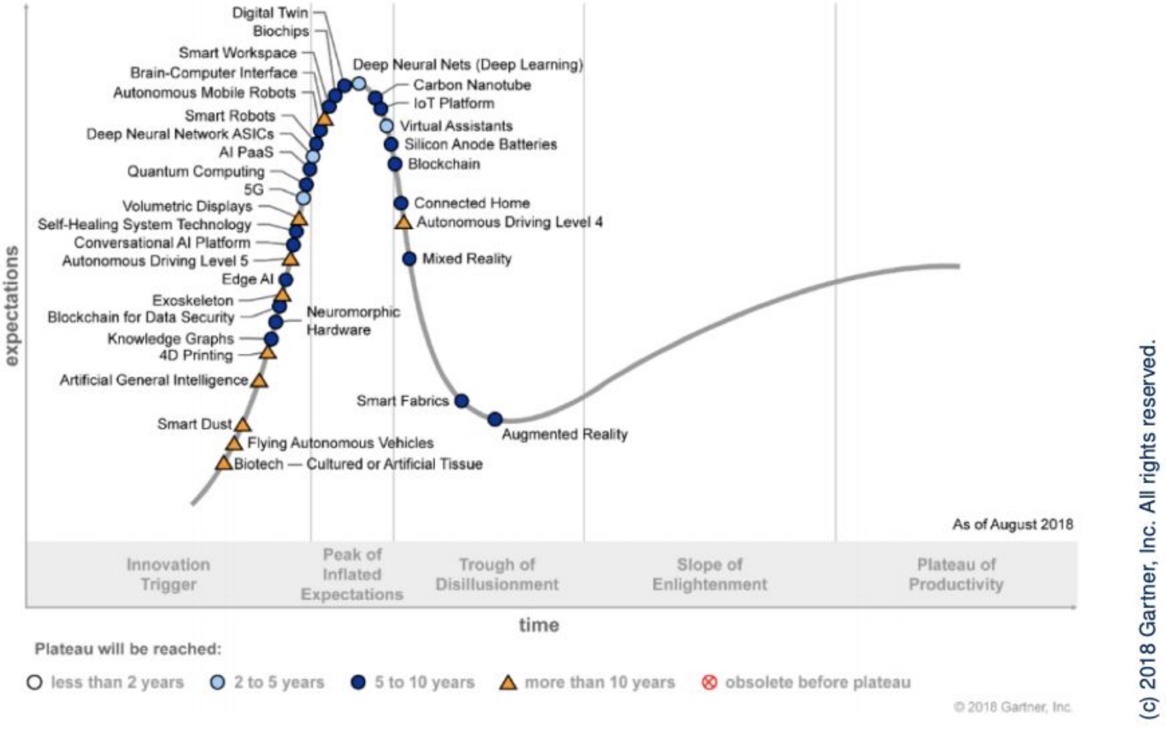


Figure 1: Knowledge Graphs Are Expected to Have High Impact During the Next 5-10 Years.

2. The Open Event Knowledge Graph (OEKG)

The Open Event Knowledge Graph (OEKG) is the one shared place for data storage and access for the Cleopatra project. It is tightly connected with further infrastructural components of the project. For instance, information and data extracted throughout the Cleopatra Knowledge Processing Pipeline such as entities, facts and their relations, provenance and context as well as user feedback populate the event-centric knowledge graph OEKG – the focal point of integration of the steps within the pipeline as well as the individual ESR projects. OEKG makes the extracted information available to the community and makes it accessible for a wide variety of applications and application domains within and beyond the Cleopatra ITN. Furthermore all ESR projects contribute with datasets to the OEKG but also benefit from the already available date in order to conduct their research work.

The Open Event Knowledge Graph has four main characteristics that should be highlighted and that we therefore discuss in more detail – multilingual, event-centric, multiple application domains, temporal data.

Multilingual datasets for cross-lingual information processing

Information about events but also about individual entities is usually available in a multitude of different languages, which is also characteristic for the diversity nature of the EU languages. Therefore, one main aspect of the OEKG is to be able to reflect this multilinguality. As a result, this enables the ESRs to advance cross-lingual information processing by alignment, validation and contextualisation of event-centric textual and visual information spread across heterogeneous multilingual sources. There new analytic methods will enable us to better interpret and understand the cross-cultural differences and related culture- and language-specific information representation in a variety of European languages, including increased support for under-resourced languages such as EU-official languages of member states that entered the EU after 2004 (e.g. bg, cz, hr, hu, pl, ro, sk, and sl). Similarly the multilingual properties of the OEKG enable us to develop novel interactive user access models to cross-lingual information, facilitating users to obtain key insights in the information presented in a foreign language and effectively interact with multilingual information at different levels, e.g. to answer questions, execute micro-tasks or perform cross-cultural event-centric analytics. Finally, based on the OEKG we are also able to develop cross-lingual and cross-cultural analytics through realizing models that describe cross-cultural information propagation in a data-driven, application-centric manner and exemplify several event-centric case studies within the politics and sports topical areas.

Event-centric data and event-based data analytics

The OEKG serves as the integral point for sharing and accessing data. In particular, existing event-centric multilingual data sets provided by the participating organisations such as news collections from EventRegistry (http://eventregistry.org/), social media collections, language resources repositories, and Web archive data collections are used and enriched throughout the individual ESR projects. Furthermore, data collections collaboratively created and enriched by ESRs are also made available through the OEKG. In this way the OEKG facilitates the advanced processing of event-centric textual and visual information on a large scale through the development of novel methods for extraction, alignment, verification, and contextualisation of multilingual information.

Multiple application domains

The OEKG aims not only to support different types of analysis but also to cover open domain as well as domain-specific data. The addressed domains include archiving, publishing, media monitoring, semantic services and journalism. These domains pose specific requirements and challenges that can only be addressed in a holistic interdisciplinary and integrated fashion, requiring expertise from several disciplines and sectors. Naturally, some of these domains will be covered in more depth than others, however, the main advantage remains – the OEKG serves as the basis for developing domain-specific approaches and solutions.

Temporal data

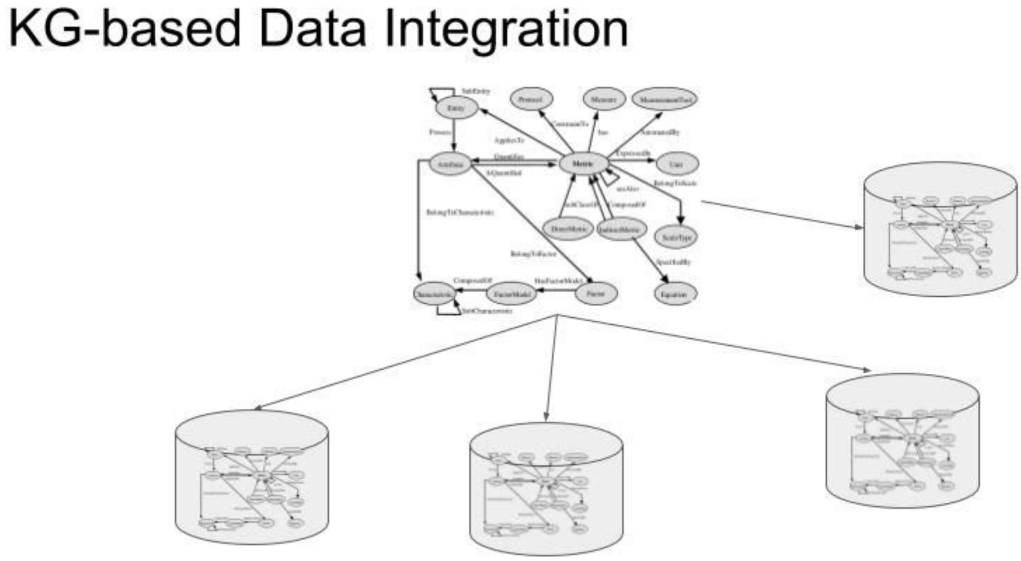
Event-centric data commonly has also a temporal aspect, which is unfortunately not represented in a lot of the available datasets. Information about similar and related events in multilingual Web can vary greatly in terms of vocabulary use, granularity, temporal resolution and level of details. The goal is to include temporal data as part of the OEKG in order to be able to develop methods that extract event-centric facts from multilingual sources and to align these facts through establishing semantic and temporal relations between the facts and their multilingual context. An accurate cross-lingual fact alignment requires better understanding of temporal event development and cross-lingual information propagation.

In the following we describe how the OEKG is realized in terms of architecture and data integration.

3. Architecture and Data Integration Approach

Designing a good data architecture and appropriate approaches for supporting diverse data integration are two tasks that are key for providing high-quality knowledge graphs. Here we describe the solutions that we apply for realizing the OEKG.

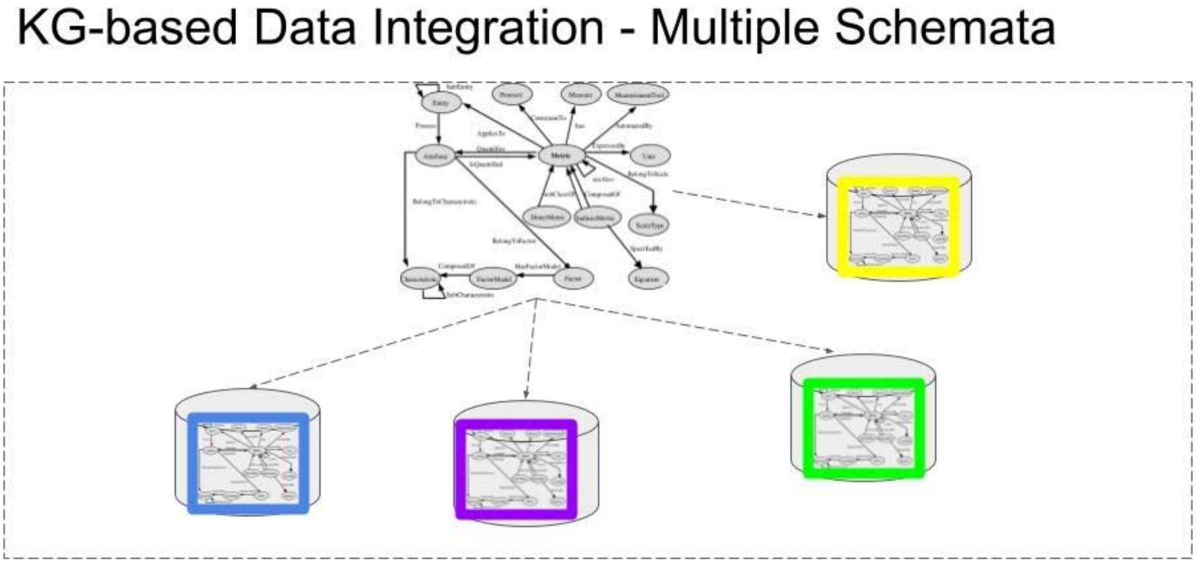
As depicted in Figure 2., the easiest way for developing a KG and providing data integration is based on using a single shared model. In this case we also talk about a top-down approach – a single model for describing the data is developed and this model is propagated and used for all datasets that are to participate in the KG. In this case there are no mismatches between the used concepts, no need for entity linking or mapping. All the used data is described based on the one and same model. This approach also usually results in very clean and consistent solutions. Unfortunately it cannot be applied for realizing the OEKG for a multitude of reasons. First of all, we aim to incorporate already existing datasets, which already have predefined schemas. The new datasets, which are developed by the ESRs are also quite heterogeneous and it would be hard to make them all stick to the same model. Therefore do not apply this very basic approach to data modelling and integration.



KG-based Data Integration

Figure 2: Knowledge Graphs Integration Based on a Single Model

Another approach for integrating multiple datasets is visualized in Figure 3. Here each of the available datasets has its own model (shown in blue, purple, green and yellow), however, there is also a shared model (visualed is gray at the top), which is used for the data integration. This approach allows the original datasets to remain more or less unmodified, however, it also requires the definition of mappings between the individual models and the one shared model. When it comes to data integration solutions, this approach is very frequently applied since it offers both flexibility and backward compatibility.



KG-based Data Integration - Multiple Schemata

Figure 3: Knowledge Graphs Integration Based on Mapping Multiple Models

Finally, a third approach for data integration is visualized in Figure 4. In addition to having individual models for each datasets, in this case there are also “intermediary” models that support the mapping to the one shared model. This approach is applied when the individual datasets are very heterogeneous and a direct mapping to the shared model is very difficult to achieve. Another common application use case is with legacy data systems, where the data and the models have evolved and continue to evolve over time. By introducing an intermediary model, this aspect of allowing for independent evolution of the data, can still be preserved. This approach obviously provides a lot of flexibility but at the same time adds complexity and might result in inconsistencies or contradictions between the intermediary models.

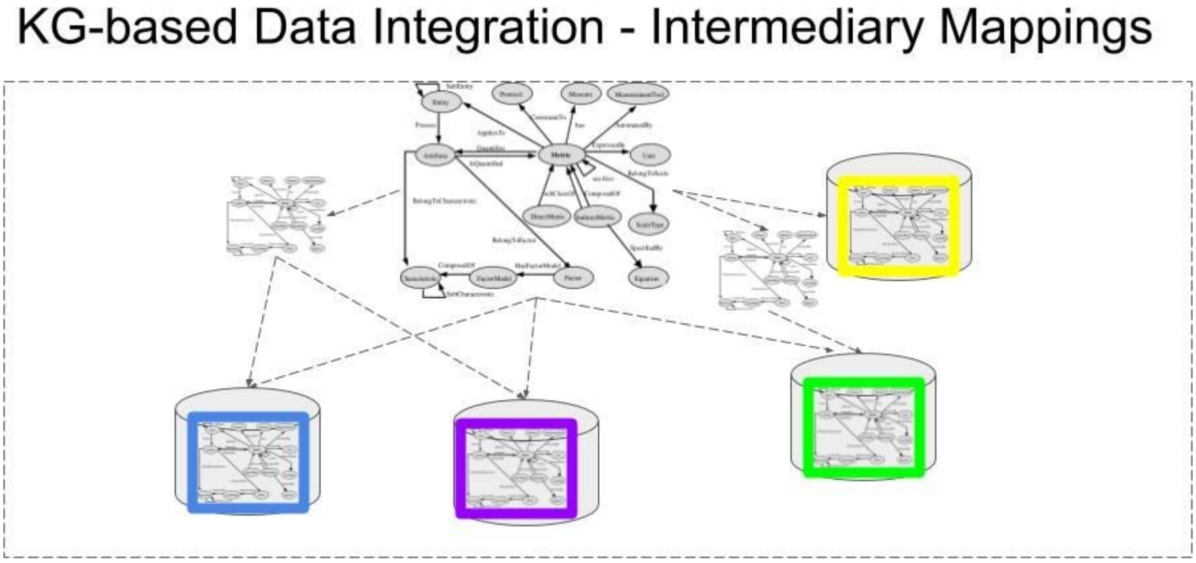

KG-based Data Integration - Intermediary Mappings

Figure 4: Knowledge Graphs Integration Based on Intermediary Models

The approach for data integration that we follow for the OEKG is visualized in Figure 5. It uses the EventKG (http://eventkg.l3s.uni-hannover.de/) model as the integration model. Based on this, individual datasets are integrated, by creating mappings between their individual model and the EventKG model. This solution allows for a lot of flexibility and also provides the basis for satisfying the four main knowledge graph characteristics listed in the previous section. The first version of the OEKG is described in detail in Section 5.

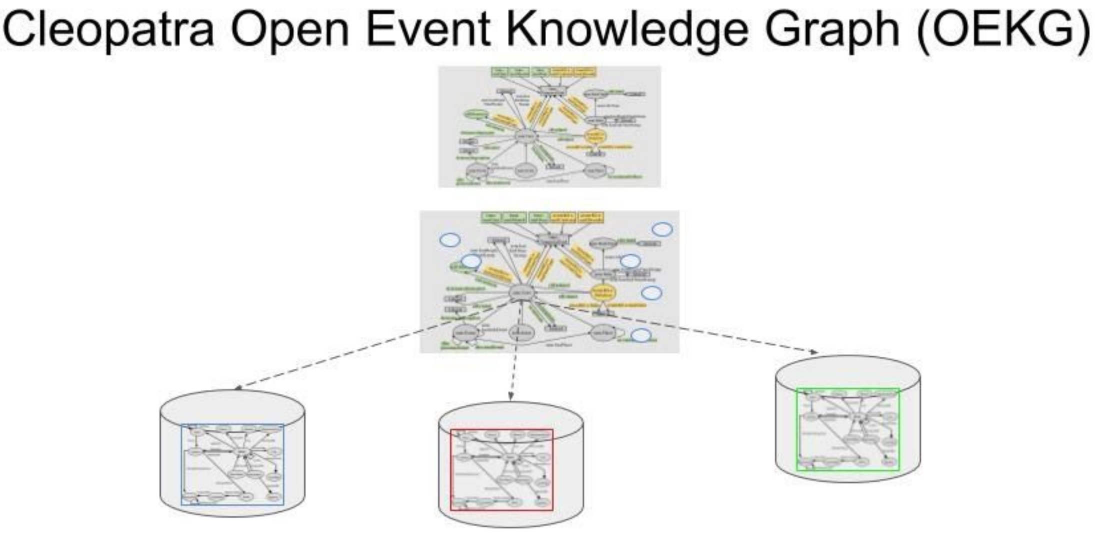

Cleopatra Open Event Knowledge Graph (OEKG)

Figure 5: Knowledge Graphs Integration Based for the OEKG

In the following section we provide an overview and a short description of all the datasets that are available for use in the Cleopatra project. These include already published datasets but also datasets that are being currently developed within the context of the research work of the ESRs.

4. Contributions of the Individual Research Projects

In this section we provide detailed description of each of the already available and newly developed Cleopatra data sets.

Partner organization	LUH
Name of the dataset	EventKG
Description of the dataset	The EventKG is a multilingual resource incorporating event-centric information extracted from several large-scale knowledge graphs such as Wikidata, DBpedia and YAGO, as well as less structured sources such as the Wikipedia Current Events Portal and Wikipedia event lists in 15 languages. The EventKG is an extensible event-centric resource modeled in RDF. It relies on Open Data and best practices to make event data spread across different sources available through a common representation and reusable for a variety of novel algorithms and real-world applications.
Multilingual (which languages)	English, German, French, Italian, Portuguese, Russian, Spanish, Italian, Dutch, Polish, Croatian, Bulgarian, Norwegian (Bokmål), Romanian and Slovene
URL	http://eventkg.l3s.uni-hannover.de/
Dataformat (RDF, JSON, XML, text)	RDF (.nq, .ttl)
Dataset size	~ 150GB
Technical requirements (repository, libraries, ...)	SPARQL
Licensing	Creative Commons Attribution Share Alike 4.0 International
Documentation	https://github.com/sgottsch/eventkg
Further details	Publications: Simon Gottschalk and Elena Demidova. EventKG - the Hub of Event Knowledge on the Web - and Biographical Timeline Generation. Semantic Web Journal. In press.

	Simon Gottschalk and Elena Demidova. EventKG: A Multilingual Event-Centric Temporal Knowledge Graph. In Proceedings of the Extended Semantic Web Conference (ESWC 2018). SPARQL endpoint: http://eventkginterface.l3s.uni-hannover.de/sparql Example application: http://eventkg-timeline.l3s.uni-hannover.de/

Partner organization	LUH
Name of the dataset	Event-QA
Description of the dataset	Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. Event-QA dataset contains 1,000 semantic queries and the corresponding verbalisations for EventKG
Multilingual (which languages)	English, German, Portuguese
URL	http://eventcqa.l3s.uni-hannover.de/ https://doi.org/10.5281/zenodo.3568387
Dataformat (RDF, JSON, XML, text)	JSON
Dataset size	1,000 queries
Technical requirements (repository, libraries, ...)	SPARQL
Licensing	Creative Commons Attribution Share Alike 4.0 International
Documentation	
Further details	Cite as: “Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs” Tarcisio Souza Costa; Simon Gottschalk; Elena Demidova http://eventcqa.l3s.uni-hannover.de/ https://github.com/tarcisiosouza/Event-QA

Partner organization	LUH
Name of the dataset	The German Web corpus
Description of the	The German Web corpus covers all Web pages from the .de

dataset	top-level domain as captured by the Internet Archive from 1996 to 2013, the HTML portion (~30TB) with 4.05 billion captures of 1 billion URLs. Overall size is ~80TB and also includes English content. From this corpus, a collection of German news sites was created based on a set of 400 domains of German news websites.
Multilingual (which languages)	German (primarily), English
URL	Available only at LUH, on site
Dataformat (RDF, JSON, XML, text)	WARC, JSON
Dataset size	~80TB German news collection: 4.3TB (32,794,626 captures)
Technical requirements (repository, libraries, ...)	Hadoop cluster, ElasticSearch
Licensing	Research only
Documentation	http://alexandria-project.eu/datasets/german-and-uk-web-archive/ German news collection: https://github.com/tarcisiosouza/elastic-client-api
Further details	-

Partner organization	UBO
Name of the dataset	**FactBench**
Description of the dataset	*FactBench* is a multilingual benchmark for the evaluation of fact validation algorithms. All facts in *FactBench* are scoped with a timespan in which they were true, enableing the validation of temporal relation extraction algorithms. *FactBench*currently supports english, german and french. You can get the current release here.
Multilingual (which languages)	yes
URL	https://github.com/DeFacto/FactBench/tree/master/core
Dataformat (RDF, JSON, XML, text)	RDF models
Dataset size	1500 correct statements, 780 negative examples

Technical requirements (repository, libraries, ...)	SPARQL or MQL
Licensing	The MIT License (MIT)
Documentation	https://github.com/DeFacto/FactBench
Further details	Used by DeFacto

Partner organization	UBO
Name of the dataset	VQuAnDa: Verbalization QUestionANswering DAtaset
Description of the dataset	VQuAnDa is an answer verbalization dataset that is based on a commonly used large-scale Question Answering dataset – LC-QuAD. It contains 5,000 questions, the corresponding SPARQL query, and the verbalized answer. The target knowledge base is DBpedia, specifically the April 2016 version.
Multilingual (which languages)	No (English)
URL	https://figshare.com/projects/VQuAnDa/72488
Dataformat (RDF, JSON, XML, text)	JSON
Dataset size	5k samples (question, SPARQL query, answer verbalization)
Technical requirements (repository, libraries, ...)	SPARQL
Licensing	Attribution 4.0 International (CC BY 4.0)
Documentation	http://vquanda.sda.tech/
Further details	

Partner organization	FCT-FCCN
Name of the dataset	**Arquivo.pt web archive**
Description of the dataset	Arquivo.pt is a research infrastructure that preserves content written in several languages broadly interesting to the Portuguese community and related to research and

	education in general. Arquivo.pt has been developing special web collections about international events such as European Elections, online news, Wikipedia or the celebration of the 100 years of World War. Arquivo.pt also collected and preserved 50.4 million Web files related to R&D activities funded by the EU since 1994 (FP4 to FP7). All the outputs from this study were made publicly available and we believe they constitute a unique and precious resource for research activities in all fields of knowledge. Arquivo.pt provides access to its collection of historical web data through a public web user interface or an API that enables the refinement of queries (e.g. by special collection). ESRs can also have access to Arquivo.pt Big Data Analytics, based on Hadoop, to perform investigations that require large-scale automatic processing of large-scale web collections.
Multilingual (which languages)	Mostly in Portuguese, English, French and Spanish. We don't perform language restrictions. Thus, in theory documents in all languages may be found.
URL	https://arquivo.pt https://arquivo.pt/api
Dataformat (RDF, JSON, XML, text)	JSON, XML, HTML
Dataset size	6062 million web files collected from 14 million websites stored in 336 TB (compressed format)
Technical requirements (repository, libraries, ...)	Knowledge about JSON and REST APIs
Licensing	https://sobre.arquivo.pt/en/about/terms-and-conditions/
Documentation	https://github.com/arquivo
Further details	https://sobre.arquivo.pt/en/

Partner organization	University of Southampton
Name of the dataset	Global web news feed (RSS)
Description of the dataset	Monthly collections of news articles, harvested from a seeded RSS list. Each month contains around ~30 million posts. Check for duplications is required.
Multilingual (which languages)	English
URL	https://webobservatory.soton.ac.uk/datasets/NKtKuwrMei8SFQG4H
Dataformat (RDF, JSON, XML, text)	
Dataset size	Various sizes
Technical requirements (repository, libraries, ...)	
Licensing	
Documentation	
Further details	

Partner organization	University of Southampton
Name of the dataset	Crisisnet qualitative data reports (USHAHIDI)
Description of the dataset	A Collection of 7,000+ qualitative reports collected from the Ushahidi + CrisisNet platform. These have been written and curated by first responders at major disaster events (e.g. Haiti Earthquake). Each record contains a timestamp, eventID, and message/text relating to a specific event.
Multilingual (which languages)	English
URL	https://webobservatory.soton.ac.uk/datasets/3cZxMoGEfmoMCTEA7
Dataformat (RDF, JSON, XML, text)	Text
Dataset size	Various sizes
Technical requirements (repository, libraries, ...)	

Licensing	
Documentation	
Further details	

Partner organization	TIB
Name of the dataset	Im2GPS
Description of the dataset	Im2GPS is a test set for geolocation estimation. The test set contains 237 geo-tagged photos, where 5% depict specific touristic sites and the remaining are only recognizable in a generic sense. The test set was originally crawled from Flickr.
Multilingual (which languages)	-
URL	http://graphics.cs.cmu.edu/projects/im2gps/
Dataformat (RDF, JSON, XML, text)	JPG
Dataset size	237 images, 40.8 MB
Technical requirements (repository, libraries, ...)	-
Licensing	Creative commons licenses
Documentation	-
Further details	GPS tags of Im2GPS test set must be extracted from EXIF data

Partner organization	TIB
Name of the dataset	Im2GPS3k
Description of the dataset	Im2GPS3k is a test set for geolocation estimation. The test set contains 3,000 geo-tagged images different than images in the Im2GPS benchmark. The dataset was originally collected from Flickr.
Multilingual (which languages)	-
URL	http://www.mediafire.com/file/7ht7sn78q27o9we/im2gps3ktest.zip

Dataformat (RDF, JSON, XML, text)	JPG
Dataset size	3,000 images, 479.1 MB
Technical requirements (repository, libraries, ...)	-
Licensing	Creative commons licenses
Documentation	-
Further details	GPS tags of the Im2GPS3k test set must be extracted from EXIF data

Partner organization	TIB
Name of the dataset	MP-16 dataset
Description of the dataset	The MediaEval Placing Task 2016 (MP-16) dataset is a subset of the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset and includes around five million geo-tagged images from Flickr without any restrictions. The dataset contains among photos of well known places and landmarks also ambiguous photos of, e.g., indoor environments, and food.
Multilingual (which languages)	English
URL	http://multimedia-commons.s3-website-us-west-2.amazonaws.com/?prefix=subsets/YLI-GEO/mp16/metadata/
Dataformat (RDF, JSON, XML, text)	SQL
Dataset size	4.7 M training images
Technical requirements (repository, libraries, ...)	
Licensing	Creative *commons* licenses
Documentation	
Further details	

Partner organization	TIB
Name of the dataset	Date Estimation in the Wild Dataset

Description of the dataset	Collection of Flickr images for predicting when an image has been taken. The meta information provided was gathered by the Flickr API server and covers a range from 1900 to 1999.
Multilingual (which languages)	English
URL	https://doi.org/10.22000/0001abcde https://github.com/TIB-Visual-Analytics/DEW-Downloader
Dataformat (RDF, JSON, XML, text)	JPG, CSV
Dataset size	1,029,710 images
Technical requirements (repository, libraries, ...)	Python
Licensing	Meta: CC BY 4.0 Attribution, Images: meta.csv
Documentation	This package contains: - Meta information for 1,029,710 images (meta.csv) - Each line in meta.csv represents: - img_id: Unique Flickr image id in the dataset - GT: Ground truth acquisition year - date_taken: The time at which the photo has taken according to Flickr - date_granularity: Accuracy to which we know the date to be true according to Flickr (https://www.flickr.com/services/api/misc.dates.html) - url: Weblink for the image. - username: Flickr username of the author - title: Image title on Flickr - licence: Image license according to Flickr - licence_url: Weblink for the license (if available) - List of images for test (test_images_1120_shuffeled.csv) and validation (validation_images_8495.csv) - Each line in test_images_1120_shuffeled.csv and validation_images_8495.csv represents: - GT: Ground truth acquisition year - img_id: Unique Flickr image id in the dataset - Instructions to download the images (download_instructions.txt):

	- Instructions to download the dataset with the source code in the following GitHub repository: https://github.com/TIB-Visual-Analytics/DEW-Downloader - Reported results (folder results_ECIR2017) in the paper - Each file predictions_* provides the predicted year of every image in the test set - Each line in predictions_* represents: - GT: Ground truth acquisition year - img_id: Unique Flickr image id in the dataset - prediction: Predicted acquisition year of the approach * or h_k: Predicted acquisition year of human annotator k - Each file results_* provides the reported results in the paper - Each line in results_* represents: - period: The results of the specific period - ME: Absolute mean error - EE_n: Percentage of images with an absolute estimation error of at most n years
Further details	Publication: E. Müller, M. Springstein, R. Ewerth: "When was this picture taken?" – Image Date Estimation in the Wild In: Proceedings of 39th European Conference on Information Retrieval (ECIR), Aberdeen, UK, 2017, 619-625. https://link.springer.com/chapter/10.1007/978-3-319-56608-5_57

Partner organization	TIB
Name of the dataset	Semantic Image-Text-Classes
Description of the dataset	This dataset is comprised of image-text pairs of eight different semantic image-text classes. Pairs of images and text can be distinguished into these classes by observing their purpose and

	classifying their interplay in the process of conveying information. The dataset consists of 224,856 (automatically labeled) image-text pairs for training and 800 pairs with human verified labels for testing.
Multilingual (which languages)	English
URL	https://doi.org/10.25835/0010577
Dataformat (RDF, JSON, XML, text)	PNG and JSON
Dataset size	225,656 image-text pairs, 45.3 GB
Technical requirements (repository, libraries, ...)	-
Licensing	Creative Commons Attribution-NonCommercial 3.0
Documentation	-
Further details	Otto, C., Springstein, M., Anand, A., Ewerth, R., “Understanding, Categorizing and Predicting Semantic Image-Text Relations”, ACM International Conference on Multimedia Retrieval (ICMR), Ottawa, Canada, 2019.

Partner organization	FFZG, KCL, LUH
Name of the dataset	UNER (Universal Named Entity Recognition)
Description of the dataset	The dataset is composed of parallel corpora based on the content published on the SETimes.com news portal which (news and views from Southeast Europe), annotated in terms of events as defined in the ACE 2005 corpus and named entities following a new classification hierarchy composed of 3 levels: 1st level: 8 supertypes 2nd level: 47 types 3rd level: 69 subtypes
Multilingual (which languages)	Albanian, Bulgarian, Bosnian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish.
URL	TBD
Dataformat (RDF, JSON, XML, text)	XML (BIO Index based)
Dataset size	200k sentences for each language.

Technical requirements (repository, libraries, ...)	TBD
Licensing	CC-BY-SA
Documentation	Under development.
Further details	Database being developed by using pre-annotation with automatic tools of the English corpus, followed by a correction step via crowdsourcing and, finally, automatically propagated to other languages. SETimes dataset: http://nlp.ffzg.hr/resources/corpora/setimes/

Partner organization(s)	UvA
Involved ESRs	ESR 13 (Anna Jørgensen)
Name of the dataset	“2019-20 coronavirus outbreak” on Wikipedia
Description of the dataset (2-3 sentences)	The data set contains the full edit histories of the “2019-20 coronavirus outbreak” pages from 70 language versions on Wikipedia. The data set is highly multilingual containing both a wide variety of alphabets and language families, as well as language sizes (from Chinese to Scots). It is also highly multimodal: - core data: content, images, links, table of content, urls - Metadata: image captions, article categorization, reference types, url countries, user ID
Multilingual (which languages)	af ar az bcl be bg bn br ca cdo cs cv cy da de el en eo

	es et eu fa fi fr ga hak he hi ht hu hy id is it ja ka kk ko ku lij lmo lt lv mr ms my nl nn pl pt ro ru sah sc sco sq sr sv sw ta th tl tr ug uk ur vec vi

	wuu zh
URL	
Dataformat (RDF, JSON, XML, text)	JSON - Text - IP addresses - Images: links to commons.wikimedia.org
Dataset size	4,37 GB
Technical requirements (repository, libraries, ...)	None
Licensing	Creative Commons Public Licence
Documentation	Here (will be migrated to data storage soon)
Publications	Forthcoming
Further details	“2019-20 coronavirus outbreak” on Wikipedia is due for release in ultimo March 2020

Partner organization(s)	LUH
Involved ESRs	ESR2 (Sara Abdollahi)
Name of the dataset	EventKG+Click
Description of the dataset (2-3 sentences)	EventKG+Click is a novel cross-lingual dataset that reflects the language-specific relevance of events and their relations. This dataset aims to provide a reference source to train and evaluate novel models for event-centric cross-lingual user interaction, with a particular focus on the models supported by knowledge graphs. EventKG+Click consists of two subsets: 1. EventKG+Click_event which contains relevance scores, location-closeness, recency and Wikipedia link count factors for more than 4 thousand events; and 2. EventKG+Click_relation with nearly 10 thousand event-centric click-through pairs, and their language specific number of clicks, relation relevance and co-mentions of the relation which is the number of sentences in whole Wikipedia language editions that mentions both the source and target.
Multilingual (which languages)	English, German, Russian

URL	https://github.com/saraabdollahi/EventKG-Click
Dataformat (RDF, JSON, XML, text)	TSV
Dataset size	3 MB in total 4113 events in EventKG+Click_event 9119 event-centric click-through pairs in EventKG+Click_relation
Technical requirements (repository, libraries, ...)	
Licensing	CC BY-SA 4.0
Documentation	
Publications	Sara Abdollahi, Simon Gottschalk, and Elena Demidova. "EventKG+Click: A Dataset of Language-specific Event-centric User Interaction Traces." CLEOPATRA – 1st International Workshop on Cross-lingual Event-centric Open Analytics, 2020.
Further details	

Partner organization(s)	UBO, TIB, JSI
Involved ESRs	ESR 5 (Jason Armitage), ESR 6 (Endri Kacupaj), ESR 8 (Golsa Tahmasebzadeh), ESR 12 (Swati)
Name of the dataset	Wiki-MLM
Description of the dataset (2-3 sentences)	Wiki-MLM is a processed data extraction from Wikipedia for multilingual and multimodal tasks. The primary aim is to train and evaluate systems designed to perform multiple tasks over diverse data.
Multilingual (which languages)	English, French, German
URL	
Dataformat (RDF, JSON, XML, text)	Text, geo-coordinates, triples - JSON Images - PNG
Dataset size	≈150k samples (four modalities per sample)
Technical requirements (repository, libraries, ...)	None
Licensing	Creative Commons Public Licence

Documentation	
Publications	Paper due in April 2020
Further details	Wiki-MLM is due for first release in April 2020

The Open Event KG — Data needs and Data Contributions of Each of the ESRs

1. **ESR 1 Fact extraction and cross-lingual alignment**

ESR1: Tin Kuculo	**Fact extraction and cross-lingual alignment**
Data needs	● Multilingual news corpora ● Multilingual event knowledge graph(s) ● Training dataset for extraction and alignment
Available Datasets that can be used	EventRegistry, EventKG
Data results	● KG enrichment: events, facts, relations ● Training dataset for fact extraction and alignment
Possible ESR collaborations	ESR5, ESR7, ESR6, ESR8
Secondments data	

2. **ESR 2 Interactive user access models to cross-lingual information**

ESR2 Sara Abdollahi	**Interactive user access models to cross-lingual information**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

3. **ESR 3 Crowd quality and training in hybrid multilingual information processing and analytics**

ESR3	**Crowd quality and training in hybrid multilingual information processing and analytics**

Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

4. **ESR 4 Incentives design for hybrid multilingual information processing and analytics**

ESR4	**Incentives design for hybrid multilingual information processing and analytics**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

5. **ESR 5 Fact validation across multilingual text corpora**

ESR5 Jason Armitage	**Fact validation across multilingual text corpora**
Data needs	Multimodel dataset + Negative examples, Unstructured / Natural Language
Available Datasets that can be used	EventRegistry, EventKG, Multilingual dataset, LC-QuAD
Data results	
Possible ESR collaborations	ESR1, ESR8, ESR7, ESR6
Secondments data	

6. **ESR 6 Interactive multilingual question answering**

ESR6 Endri Kacupaj	**Interactive multilingual question answering**
Data needs	Multimodel dataset + Negative examples, Unstructured / Natural Language

Available Datasets that can be used	EventRegistry, EventKG, Multilingual dataset, LC-QuAD
Data results	
Possible ESR collaborations	ESR1
Secondments data	

7. **ESR 7 Relations of textual and visual information**

ESR7 Gullal Singh Cheema	**Relations of textual and visual information**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

8. **ESR 8 Contextualisation of images in multilingual sources**

ESR8 Golsa Tahmasebzadeh	**Contextualisation of images in multilingual sources**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

9. **ESR 9 National and transnational media coverage of European parliamentary elections**

ESR9 Daniela Major	**National and transnational media coverage of European parliamentary elections**
Data needs	
Available Datasets that can be used	
Data results	

Possible ESR collaborations	
Secondments data	

10. ESR 10 Nationalism, internationalism and sporting identity: the London and Rio Olympics

ESR10 Caio Castro Mello	**Nationalism, internationalism and sporting identity: the London and Rio Olympics**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

11. ESR 11 Information propagation with barriers

ESR11	**Information propagation with barriers**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

12. ESR 12 Cross-lingual news reporting bias

ESR12	**Cross-lingual news reporting bias**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

13. ESR 13 Multilingual Wikipedia as ‘first draft of history’

ESR13 Anna Katrine Jørgensen	**Multilingual Wikipedia as ‘first draft of history’**
Data needs	All Wikipedia content containing information about certain specific events (e.g. Brexit), in Germanic languages (e.g. Dutch, Danish, Swedish, Norwegian, Faroese). Temporal and spatial information about creation and edits to pages Information about editors Repository of culture specific facts for events. News articles and temporal development of certain specific events (e.g. Brexit), in Germanic languages (e.g. Dutch, Danish, Swedish, Norwegian, Faroese). Annotated data for low-resource North Germanic languages for NLP tools.
Available Datasets that can be used	EventKG, EventRegistry, Web Archives
Data results	
Possible ESR collaborations	1, 2, 5, 14, 15
Secondments data	

14. ESR 14 NLP for under-resourced languages

ESR14 Diego Alves	**NLP for under-resourced languages**
Data needs	
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

15. ESR 15 Cross-lingual sentiment detection

ESR15 Gaurish Thakkar	**Cross-lingual sentiment detection**
Data needs	Sense Tagged Data, Parallel Corpora/Comparable Corpora(similar topic different content), Dataset with Aspects extracted along with

	their sentiments, Senti-wordnet, Machine Translation Models - From different domains (news,finance..) - Nature of data-> noisy data/social media text, maybe code-mixed data If multi-modality(text+image) is desired then dataset with (text,image)->(sentiment) is required
Available Datasets that can be used	
Data results	
Possible ESR collaborations	
Secondments data	

5. Release of the Open Event Knowledge Graph Version 1.0

The core of the Open Event Knowledge Graph (OEKG V1.0) is built on the EventKG V3.0 released on 31 March 2020 (https://zenodo.org/record/3733829). EventKG is an event-centric multilingual knowledge graph modelled in RDF. The light-weight RDF data model of EventKG/OEKG facilitates seamless integration and fusion of heterogeneous event representations and temporal relations extracted from a variety of sources and makes this information available to real-world applications through standardised RDF representation.

OEKG V1.0/EventKG V3.0 released in the Cleopatra project provides event-centric information in 15 languages extracted from several sources, including Wikipedia, Wikidata and DBpedia. Compared to the previous release of EventKG (EventKG V2.0) that included six languages, namely English, French, German, Italian, Russian and Portuguese, in this version, we have significantly extended the language coverage. The current release includes nine additional European languages, namely: Spanish, Italian, Dutch, Polish, Croatian, Bulgarian, Norwegian (Bokmål), Romanian and Slovene, some of which are under-resourced.

Cleopatra ESRs have contributed to the extraction of language-specific knowledge to populate the knowledge graph, leading to a significant increase in the number of languages as well as covered events. In total, the OEKG V1.0 contains data about 1,348,143 events, compared to less than one million in the previous EventKG release.

The following table provides an overview of the language-specific event coverage. For example, OEKG V1.0 provides an English label or description for 847,429 events, which makes English the most covered language. Within these events, 396,423 are covered in English only. This table also indicated the complementarity of the events in different languages. For example, there are approximately ten thousand events which are only covered in Slovene, including the local elections in Slovenia in 2014. For only 830 events there are labels available in all 15 currently considered languages, including, e.g. the two World Wars.

Language	**Covered Event Labels/Descriptions**	**Language-specific events**
English	847,429	396,423
French	343,232	121,646
Dutch	332,975	39,028
German	253,469	34,155
Spanish	178,257	35,529
Italian	172,936	25,552
Russian	148,392	49,257
Polish	126,498	49,006
Portuguese	88,741	20,317
Norwegian (Bokmål)	44,596	10,202
Danish	32,669	5,451
Romanian	27,688	9,394
Slovene	25,921	10,011
Bulgarian	24,681	5,741
Croatian	14,689	3,999
All	1,348,143 (all events)	830 (events with labels/descriptions in all 15 languages)

6. Planned Future Work

The next release of the Open Event Knowledge Graph is planned for month 25 (end of January 2021). Until then we will focus on continuously improving the individual datasets and the overall integration. In addition to that we will focus on facilitating and further developing the core characteristics of the OEKG, making sure to include data that is multilingual, event-centric, covers multiple application domains and has temporal aspects.

All ESRs will be contributing with the datasets that they use to the OEKG. Furthermore, in the long-term the maintenance and continuous improvement of the knowledge graph should also be opened-up to the community, which will ensure its wider use and adoption.