The InTaVia Knowledge Graph – Publishing European National Biographical and Cultural Heritage Object Data

Matthias Schlögl[1] [0000-0003-1451-0987], Jouni Tuominen[2,3] [0000-0003-4789-5676], Joonas Kesäniemi[3] [0000-0002-3770-0006], Petri Leskinen[2,3] [0000-0003-2327-6942], Victor de Boer[4] [0000-0001-9079-039X], Go Sugimoto[4] [0000-0003-2646-6784], and Joh Dokler[5] [0000-0002-8053-8198]

[1] Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Vienna, Austria
[2] University of Helsinki, Helsinki, Finland
[3] Aalto University, Helsinki, Finland
[4] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[5] ZRC SAZU, Ljubljana, Slovenia

Abstract. In this paper, we describe the InTaVia Knowledge Graph (IKG), a large Knowledge Graph containing heterogeneous multilingual data from European national biographies, connected to related cultural heritage objects. The knowledge graph was constructed in the context of the H2020 project "In/Tangible European Heritage – Visual Analysis, Curation and Communication" that aims to provide researchers and the informed public access to such biographical information. We describe the source data, data model and resulting knowledge graph as well as the pipeline components for managing and harmonizing the data. The data model used combines domain standards CIDOC CRM and Bio CRM with features to represent multiple perspectives on biographical information. In addition to the data model and data itself, the infrastructure used to harmonize and maintain the heterogeneous data is of interest not only to the Digital Humanities (DH) community but also to the Semantic Web community.

Keywords: semantic web · linked open data · biographies · cultural heritage data · data integration

1 The InTaVia project

The main aims of the European project "In/Tangible European Heritage – Visual Analysis, Curation and Communication" (InTaVia)[6] are integrating structured data from four national biographical dictionaries, enriching this data with cultural heritage objects (CHO) from reference resources and providing a web based visual analytics component that allows to gain new insights in the data. Since

[6] https://intavia.eu

these original dictionaries are established in different locations, over long periods of time using different curation strategies, the data that was to be integrated is highly heterogeneous. We therefore turned to semantic web technology and bring these datasets together into a single integrated knowledge graph, while keeping their richness in tact. This paper presents that InTaVia Knowledge Graph (IKG).

To our best knowledge this is the first attempt to harmonize structured data extracted from a set of national biographies in a knowledge graph and further enrich it using linked open data resources. In addition to the data itself we believe that the data model and the infrastructure used to harmonize the data is of interest especially to the Digital Humanities (DH) community. At the same time, the presented resource is of interest to the Semantic Web community since it brings together very rich information in a large knowledge graph. It provides a central and open resource to connect other biographical or other historical data to for a variety of applications and can serve as a benchmark dataset for (machine learning) methods and tools that deal with such kind of data.

This paper covers the IKG as well as the conversion and data harmonization infrastructure used to construct and maintain it. This serves also as a backend to the InTaVia platform, that includes a frontend. We here describe the source datasets, the ontology and the data modeling, provide an overview of the data processing pipelines and cover the Rest API created to provide the data to the InTaVia frontend.

The GitHub repositories containing all the source code, most of the data and the issues that were used to discuss and decide on problems are available via our GitHub organisation[7]. Other available resources include the beta version of the frontend application that allows to query and visualize the data[8], the API that that is consumed by the frontend, but can also be used by other applications/users[9] and the read-only SPARQL endpoint[10]. Currently, the data the ingestion process is being refactored to a GitHub driven process (please see down below for details), the repository that will contain all the datasets (including the enrichment) is also available on GitHub[11].

2 Related work

Several Knowledge Graphs in the domain of cultural heritage and digital humanities that present rich information from a specific heritage institute have been published. Examples include the Rijksmuseum Linked Data [4], the Prado knowledge Graph [1], the Smithsonian American Art Museum Linked Data [19] or the Amsterdam Museum knowledge graphs [3]. Such knowledge graphs can be reused for a variety of tasks relevant to the Semantic Web and Data science

[7] https://github.com/intavia

[8] https://intavia.acdh-dev.oeaw.ac.at

[9] https://intavia-backend.acdh-dev.oeaw.ac.at/v2/docs

[10] https://intavia-ts.acdh-dev.oeaw.ac.at

[11] https://github.com/InTaVia/source-data

communities. For example, the Amsterdam museum Dataset has been used as a benchmark for various Knowledge Graph Learning methods.

While the abovementioned knowledge graphs typically are consolidated from one source, there have also been several knowledge graphs that bring together datasets from various sources, and that keep some of the heterogeneity intact. The "Sampo" series of semantic portals all describe several aspects of (mostly Finnish) cultural heritage and history using knowledge graphs. These aspects range from culture [10], books [15], war history [12] or academic history [14]. Several of these knowledge graphs explicitly model biographical data. The Dutch Ships and Sailors graph [2] does this for several Dutch maritime historical datasets. Europeana provides access to its aggregated collections through a SPARQL endpoint [9]. Like Europeana, the InTaVia knowledge graph combines information from various, heterogeneous sources, however, the goal is to do this in a very rich datamodel, that keeps in tact the complexity of the original sources, to allow for deep scrutiny needed in the digital humanities context.

For creating the Austrian data used in InTaVia a software framework called APIS [17] was used. APIS is based on a relational database, but allows to export data created within the application in various formats (an internal JSON, TEI and CIDOC CRM based RDF). In addition to the web GUI that allows to create/update/delete the data it also provides a Rest API including an OpenAPI 3 definition that allows to easily attach external applications to the framework (e.g. Social Network Analysis tools).

In general, for publishing biographical/prosopographical data a variety of technologies have been used: TEI (such as the Slovenska biografija [7]), relational databases (such as the APIS dataset), document based databases/search indexes (such as the Neue Deutsche Biographie[12]) and RDF based systems (such as BiographyNet [8] and BiographySampo [11]) (see [18] for an overview table).

In terms of biographical knowledge graphs, we build on previous work in the BiographyNet project, where a knowledge graph for Dutch biographies was constructed, as well as the BiographySampo system for Finnish biographies. The InTaVia Knowledge Graph borrows from the BiographyNet model [16] the ability to describe multiple perspectives on persons, via the ORE-OAI Proxy model [13], which itself was borrowed from the Europeana Data Model [6]. The main part of our datamodel however is based on the domain standard CIDOC CRM [5]. Its event-centric model is very suitable to describe a variety of heterogeneous objects. The Bio CRM extension provides additional constructs and design patterns to describe biographical information, specifically towards prosopographical descriptions [20].

3 The source data

InTaVia brings together data from four national biographical dictionaries: Austria (APIS), Finland (BiographySampo), Slovenia (SBI) and the Netherlands (BiographyNet).

[12] http://www.ndb.badw-muenchen.de/ndb_aufgaben_e.htm

3.1 APIS

The APIS dataset created from the Austrian Biographic Dictionary (ÖBL)[13] contains 18 179 distinct person entities. Almost all of these person nodes contain birth and death events, even if some of them contain only inaccurate dates and some are missing relations to places. The APIS web application itself covers more persons (30 879), but only those 18 179 with full biographical information in the Österreichische Biographische Lexikon (ÖBL) were imported to the IKG. This degree of completeness and detail satisfactorily covers the meaningfulness of the source dataset.

Of the 18 332 place entities of the source dataset 7019 are included in the InTaVia triplestore, those with relations to ÖBL persons and/or institutions. Of the 3709 institutions of the source dataset 3257 are represented as CIDOC CRM *E74 Group* in the IKG. The named graph of the APIS data includes 46 577 relations linking individuals to occupational categories. APIS contains 61 577 person-event-relations. There are also about 2700 events containing relations to institutions.

The APIS data includes data created by researchers via manual annotations. While the original API provides provenance data on who created these annotations, the IKG serialization currently misses these provenance metadata. We are working on a model and a serialization to include this important information.

3.2 BiographySampo

The IKG currently contains 5833 out of ca. 7000 person entities covered in BiographySampo (BS), which is based on the National Biography of Finland and other biographical databases of the Finnish Literature Society[14], interlinked with related data repositories. People still alive, fictional characters as well as actors representing families or kins are not included in the IKG. For all these entities birth and death are recorded as corresponding events (E67 Birth and E69 Death) and are modelled with 33 944 resources in the class E33 E41 Linguistic Appellation. Furthermore, the data contains approx. 103 000 lifetime events of the actors modelled as instances of E5 Event or its subclasses.

Of the 4969 place entities of the source dataset 3889 are included in the triplestore. The source dataset also contains links to images of persons, which are represented by 3050 entities of the class E36 Visual Item. Altogether 5642 people have a link (owl:sameAs) to a corresponding Wikidata instance.

Additionally, the IKG contains currently 797 BiographySampo occupation categorizations for persons and 2745 family relationship roles.

3.3 BiographyNet

The IKG contains all 79 412 person entities from BiographyNet (BNet). This dataset is the only main source dataset, that contains potentially multiple biogrphical descriptions of the same person. While the data had previously been

[13] https://www.oeaw.ac.at/acdh/oebl

[14] https://kansallisbiografia.fi/english

converted to Linked Data, for the harmonization, the original source data from the Biography Portal of the Netherlands (BPN)[15], was re-converted. It provides 225 754 different 'proxies' (or biographical perspectives) for the 79 412 persons. Currently the data from different sources are stored in one named graph and structured with aggregations (OAI/ORE) of person descriptions, which is compatible with the proxy construction we use in IDM-RDF. The BNet dataset is currently the only one with references to various sources, because it is the only dataset which is created from different sources.

The BNet dataset contains a vast number of relations, which add great expressiveness to the dataset. That includes 277 350 relations about the participation of persons in events, 308 204 relations which connect events with places and 261 229 relations between events and their duration.

BNet data about images, graphics, source texts, occupations, gender (bn:sex), residence, education, faith, other person categorizations (bgn:StateEvents), revisions of data and events like baptism and funerals are included in the triplestore, but they are still modelled with classes from the BNet data model. This needs to be corrected in the next iteration of the data alignment process to create a useful graph. That being said, the current state and especially the size of the current BNet data is very promising. The BNet dataset is the only data until now, which contains data about the nationality of persons. However, since the selection of individuals for the BNet dataset is based on their Dutch citizenship, this data is only meaningful if we manage to aggregate other citizenship data and add it to the InTaVia triplestore.

3.4 SBI

The IKG contains 7908 of 11 660 person entities from the Slovenska biografija source (the New Slovenian Biographical Lexicon included) dataset and 7641 birth events and 6801 death events. All 7908 person entities have relations regarding their gender assignment and identifiers like names and IDs. The dataset contains 178 relations to the place where an event occurred.

We additionally enriched the SBI data by manually adding CHOs and events to 11 persons (10 selected from the richly annotated set of biographies published in the volume A of the New Slovenian Biographical Lexicon & one from the SBI). CHO types include literary works, scientific works, musical compositions, motion pictures. Events include memberships and work-related postings. These new annotations will be added to the IKG as a separate graph to distinguish them from the automatic import of the base SBI data.

We are further planning to extend the data by automatically (using machine learning methods) identified objects and places/events from the subset of the SBI data. The objects and places/events are already identified and annotated in the SBI TEI XML data format. What needs to be done is to manually review these and add the necessary information (e.g., place IDs, relations, normalised labels) that will allow us to serialize this information to RDF.

[15] http://www.biografischportaal.nl/en/

4 Infrastructure

The InTaVia infrastructure is hosted at the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) at the Austrian Academy of Sciences. It consists of the following components:

- A Blazegraph[16] triplestore in quad mode as the database backend
- A FastAPI[17] backed Rest API that consumes the triplestore and delivers JSON to the frontend
- A React based frontend that allows to search the IKG and perform visual analytics on the data
- A Prefect[18] (v1) based ETL (Extract, Transform and Load) framework that is used to run enrichment and/or curation jobs on the IKG and pull in data from the original sources (e.g APIS)
- A ResearchSpace[19] instance, used to manually inspect the IKG and find errors

The API, the frontend and the Prefect flows are developed on GitHub and published under the MIT license. GitHub actions are used to deploy the services directly to the ACDH-CH Kubernetes[20] cluster.

5 The ontology (IDM-RDF)

To combine cultural heritage and biographical data in IDM-RDF core, a widely used, established, flexible data modeling tool was required that would cover the large field of cultural heritage and prosopography while allowing detailed data to be provided on these topics. It was decided to model the IDM-RDF according to the CIDOC CRM v7.1.1 implemented in RDFS.

The decision to use CIDOC CRM as a basis for IDM has been taken for the following reasons:

- The CIDOC CRM is an ISO standard for the exchange of cultural heritage data since 2006.
- The CIDOC CRM has been under active development since the 1990s.
- The CIDOC CRM is widely adopted by the Digital Humanities community.
- The CIDOC CRM is already in use by two out of four biographical data providers (ACDH-CH, Aalto University).
- The CIDOC CRM covers the requirements of the Dublin Core Metadata Initiative data model which was adapted by the Europeana Data Model.

[16] https://blazegraph.com
[17] https://fastapi.tiangolo.com
[18] https://www.prefect.io
[19] https://researchspace.org
[20] https://kubernetes.io

– The CIDOC CRM provides ”definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation and of general interest for the querying and exploration of such data”[21].

Even though IDM-RDF is event-based (like CIDOC CRM), there is a group of data that is usually collected untemporalized in biographical lexica. These include, for example, nationality, gender, social relations, and occupations. To address this challenge, which is typical for prosopography[22], CIDOC CRM in IDM-RDF is complemented by the Bio CRM extension[23]. Bio CRM is an RDF-based CIDOC CRM extension for representing roles of actors and things in events and to allow the implementation of the untemporalized roles, such as gender, nationality, and occupation. The Bio CRM classes and properties are used as superclasses and superproperties of equivalent IDM-RDF classes and properties to allow the adaptation in the context of the IDM-RDF, which makes some changes of domains and ranges necessary.

The modeling of the CHO data was done after detailed consideration of the Europeana Data Model (EDM) and exemplary queries of the Europeana data via the official Europeana SPARQL endpoint[24] to include information on the actual available Europeana data in the modeling decisions.

The IDM-RDF data model is published on GitHub[25], with archival on Zenodo[26].

6 API and JSON schema (IDM-JSON)

During the initial project phase it was decided to create a Rest API in front of the SPARQL endpoint. Some of the most important factors for this decision were:

– Rest APIs delivering JSON are a more widely adopted than SPARQL endpoints. Even within the academic field we foresee much more people using the Rest endpoint than the SPARQL endpoint.
– Rest APIs allow to restrict queries more easily. Having a Rest API in place allows us to put the SPARQL endpoint behind authentication while still maintaining access to the data, in case the SPARQL endpoint goes down on a regular basis (due to expensive queries).

[21] See ”What is the Cidoc CRM?”, https://www.cidoc-crm.org

[22] E.g. the Swiss Art Research Infrastructure CIDOC CRM extension by George Bruseker and Nicola Carboni has an untemporalized *SRP2 had occupation* property, https://docs.swissartresearch.net/schema/#SRP2_had_occupation

[23] See Bio CRM Schema, http://ldf.fi/schema/bioc/

[24] http://sparql.europeana.eu

[25] https://github.com/InTaVia/idm-rdf

[26] https://doi.org/10.5281/zenodo.5534542

- Rest APIs are the most common middleware between databases and the frontend. Providing a Rest API including an OpenAPI 3 spec of query parameters, routes and return shapes allowed for an easier frontend development.
- The Rest API is meant to reduce some of the complexity of the data.

The API – while still in active development – is already available[27]. It is divided in *Entities*, *Events*, *Vocabularies* and *Statistics* endpoints. All endpoints have a *search*, a *get by id* and a *bulk retrieve* route. The former allows for searching and includes – depending on the endpoint – various query parameters, the *get by id* endpoints allow to retrieve a single object when the id is known and the latter allows for retrieval of several objects when the respective ids are known. Endpoints that return arrays of objects feature pagination and setting the page size.

IDM-JSON, the JSON schema delivered via the API, is separated in attributes of entities (such as gender, or longitude and latitude) and an array of temporalized events that again include an array of participating actors/entities.

6.1 The API

The API is built using FastAPI, a Python framework for building Rest APIs. One of the core advantages of FastAPI over other frameworks is its rigorous use of typing. The Pydantic library[28] is used for validating requests as well as returned data.

The InTaVia API uses Jinja templating language[29] to create the SPARQL queries posted against the triplestore. This allows to use inheritance – e.g. all queries use the same template to limit the query to a selected list of datasets – and ingest the query parameters in the SPARQL query. For creating a nested JSON out of the flat JSON returned by the SPARQL endpoint a custom Pydantic base class was created.

```
class Entity(IntaViaBackendBaseModel):
    id: str = Field(...,
        rdfconfig=FieldConfigurationRDF(path="entity",
        anchor=True, encode_function=pp_base64))
    label: InternationalizedLabel | None = Field(
        None,
        rdfconfig=FieldConfigurationRDF(
            path="entityLabel",
            default_dict_key="default")
    )
    kind: EntityType = Field(EntityType.Person,
        rdfconfig=FieldConfigurationRDF(
```

[27] https://intavia-backend.acdh-dev.oeaw.ac.at/v2/docs; The source code is available in the repository https://github.com/InTaVia/InTaVia-Backend.

[28] https://pydantic.dev

[29] https://jinja.palletsprojects.com

```
            path=”entityTypeLabel”)
        )
    ...
    relations: list[EntityEventRelation] = []
```

Listing 1.1. InTaVia API class definition example

The example above shows a selection of the class definition of the entity. The Pydantic base class developed for SPARQL JSON allows to set some configurations – such as whether to fail on validation errors – and overrides the __init__ method. When data is loaded in the class, it runs through the field definitions and uses the *FieldConfigurationRDF* class to put the data in the correct shape.

The *FieldConfigurationRDF* class allows to set the path, the key needed to retrieve the data from the JSON to populate the field, whether this key is an anchor element and pre- and postprocessing functions to apply to the data. If the key is set as an anchor element, the value of that key is deemed as unique and all rows containing that value are nested beneath the element. In conjunction with the class inheritance and the possibility to define Pydantic classes as fields – super classes only pass on the data that is needed by fields and/or child classes – this allows to build arbitrary deeply nested JSON objects out of the flat SPARQL JSON.

7 The pipeline components

For serializing the datasets, and for further enrichment and curation tasks it was decided to set up an ETL framework[30]. After some preliminary experiments with Apache Airflow[31], the consortium decided to go with Prefect. Prefect is a relatively new open-source project with an integral commercial connection. Compared to Apache Airflow, basic workflows with Prefect are easy and intuitive to create and many helpful extensions are available.

7.1 Entity reconciliation process

For reconciling entities (persons, places) in the four biographical source datasets, we have implemented an entity ID enrichment process which queries the Wikidata SPARQL endpoint for additional external entity ID’s. The process utilizes existing same-as mappings in the source datasets, that connect entities to external entity ID’s. For example, BiographySampo includes mappings (owl:sameAs) to Wikidata QID’s, which can be used for getting equivalent Integrated Authority File (GND) ID’s. Similarly, for APIS, Wikidata QID’s can be fetched based on the GND ID’s included in the source data. The IKG is enriched with these

[30] The workflows used for creating the stable IKG are based on Prefect v1 and are published in https://github.com/InTaVia/prefect-flows. The workflows for the improved GitHub-based ingestion (see 7.5) are published in https://github.com/InTaVia/prefect2-flows

[31] https://airflow.apache.org

external ID's, and the entities in, e.g., BiographySampo and APIS are reconciled (attached to the same *Provided_Entity* instance) based on the Wikidata-GND mapping.

7.2 Enrich the interperson relations from Wikidata and Getty ULAN

To enrich the number of interperson relations in IKG additional relations were extracted from Wikidata. Out of the total 58 868 IKG actors with Wikidata links approx. 10 000 had altogether approx. 18 300 links to other actors in IKG, only the relations between IKG actors were chosen. The interperson relations were related of education (student, teacher, supervisor), genealogical (parent, child, spouse, other relatives), or career-related (co-worker, influencer). Notice that e.g., family relations already might be available in some data sets like BS. The data model follows the Bio CRM schema, and the resulting data is currently added to graph <http://intavia.eu/wikidata-relation-enrichment> in InTaVia triplestore. Similarly, the interperson relations extracted from Getty Union List of Artists' Names will be added to a graph <http://intavia.eu/ulan-relation-enrichment>.

7.3 Enrich the IKG with cultural heritage objects from Wikidata

The enrichment of the IKG with CHOs from Wikidata is based on a rather simple systematic. After setting up the connection to the triplestore it uses federated queries against Wikidata to pull in the needed data. To not run into timeouts it uses a configurable while loop to iterate through all persons with Wikidata identifier in the IKG. In addition it makes use of the *prefect.task* failure/re-try mechanism to facilitate for other timeouts/problems such as network outages, hourly computation time limit etc.

Currently the enrichment flow queries for any objects connected to the person via *wdt:P170* (creator). It uses the object titles in our four languages (German, Slovene, Finnish and Dutch) and English, the inception dates and the place of creation to create a basic CHO entity connected to the person in the source-graph via a creation event and pushes that in a dedicated named graph. We are currently working on adding more types of cultural heritage objects, such as literature, and add more metadata, such as material used, to the graph.

7.4 Enrich the IKG with cultural heritage objects from Europeana

In the project plan Europeana was foreseen as the main source for cultural heritage data. However, during an initial evaluation phase, it turned out that Wikidata does have nearly as much CHOs for our data sources as Europeana has, while being much easier to query and include in the IKG. Due to these problems – mainly caused by the Europeana SPARQL endpoint – pulling in Europeana data is currently not done via SPARQL, but by batch downloading files via the Europeana API, generating a Turtle file on a local machine and uploading it to the IKG.

7.5 Improved ingestion process via GitHub

While the Prefect pipelines are running smoothly, allow to keep a centralized infrastructure for updating data and store all logs of these update process in one database, they are also not accessible for anyone outside our institutions.

Therefore we decided to move the ingestion workflow to GitHub. While the stable IKG was still populated by manual ingestions and/or Prefect pipelines, the infrastructure for the GitHub based ingestion is already in place and a development triplestore[32] is populated using the new workflow[33]. The source-data repository[34] has been setup and it is configured to allow for the following workflow (Fig. 1 has an overview of the ingestion systematic).

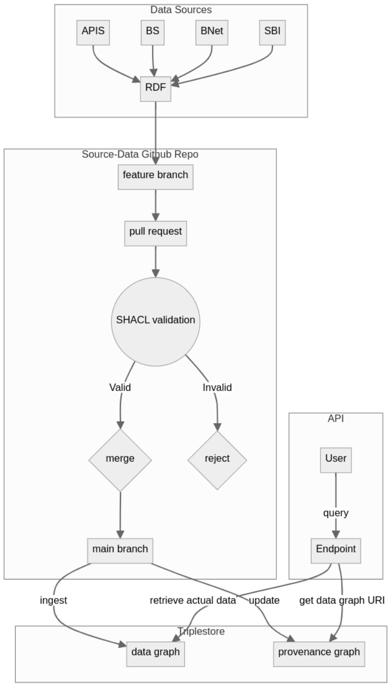

Fig. 1. Systematic of the data ingestion process via GitHub

The main branch is protected and allows for merging pull requests (PR) from other branches only. When a PR is created and contains a new/updated Turtle file in the *datasets* directory a validation using SHACL constraints[35] kept in the

[32] https://intavia-ts.acdh-ch-dev.oeaw.ac.at

[33] Please note that not all Prefect pipeline components have been ported yet to the new workflow.

[34] https://github.com/InTaVia/source-data

[35] https://www.w3.org/TR/shacl/

source-data repo is run. It succeeds only if the file contains no violations of the SHACL shapes.

PRs can only be merged on a review and a successful run of the validation action. When the PR is merged to the main branch the ingestion action is started. It posts the Turtle file against the SPARQL endpoint using a new named graph (containing the commit hash). In a second step it uses a SPARQL query to update the *valid_from*, *valid_to* and *latest_for* for the dataset in the provenance graph accordingly.

This brings several improvements compared to the previous method:

- Validation is run directly in GitHub actions, which secures valid data while releasing contributing researchers from running scripts on their workstations.
- The public GitHub repository allows everyone who wants to contribute data to create pull requests against the repo, while securing that the consortium – via the review process – still has control over what data is added to the IKG.
- Using Git we always have a complete history of the IKG and can recreate any previous state of the knowledge graph.
- The GitHub integration Zenodo is offering brings along an easy to use long term archival for the IKG datasets[36]
- The Git history allows us to keep persistent URIs for the provided entities through evolving versions. The entity reconciliation process (see Subsection 7.1) creates new URIs for the *Provided_Entity* instances in the InTaVia namespace to link together entities of different datasets. In the previous version these URIs were created by chance and the identifiers therefore changed on every run of the workflow. This makes caching in the frontend and citing impossible. The new workflow searches through the Git history to reuse URIs that have been assigned already in previous versions.

Our work on the IKG has shown that using GitHub (or any other Git based platform that allows for CI/CD) as a central data hub for creating complex knowledge graphs brings along several benefits that could also be of benefit to other projects using semantic web technologies.

8 The data

The InTaVia data can be queried and retrieved via the already mentioned REST API or via the SPARQL endpoint. Additionally the current versions of the smaller datasets can be found in the source-dataset-conversion[37] repository. Some of the source datasets are already available in the source-data repository[38], the rest including the enrichment named graphs are to follow in the upcoming

[36] Zenodo is taking a snapshot of the repo on every release and makes the releases citeable via DOIs. See https://doi.org/10.5281/zenodo.10290205 for the source-data dataset.

[37] https://github.com/InTaVia/source-dataset-conversion

[38] https://github.com/InTaVia/source-data

weeks. This data repository is hooked to Zenodo and archived on every new release. The IKG currently contains close to 25 000 000 triples which describe roughly 700 000 entities.

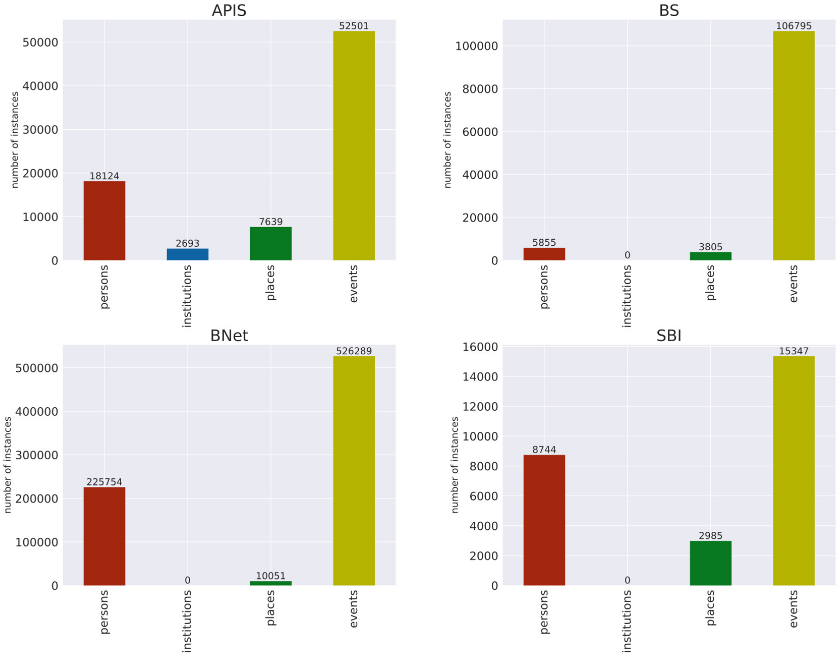

Fig. 2. Number of entities in the four biographical data sources

Fig. 2 gives an overview of the entity types across the source datasets. It clearly shows that the data is not equally distributed across the datasets. APIS for example is the only dataset that contains institutions (*crm:E74_Group*). BNet on the other hand includes by far the most person instances. BS includes – at least in relation to the overall entities – the most events. All these differences in the data can be attributed to the history of the datasets. The APIS data for example was partly manually annotated and enriched during a previous project, while the other datasets were only automatically enriched. These automatic processes are also the reason for the high number of events in relation to other entities: the automatic processes were most of the time not able – also often due to missing information in the texts – to extract all entities (e.g. the institutions in an employment event) participating in an event. Places – an entity present in all datasets – were most of the time extracted from the headline of the biography. These headlines feature the most important metadata of the depicted person (e.g. date and place of birth and death, occupations, gender, sometimes faith) and can – due to the very formal structure – be easily processed with automatic scripts.

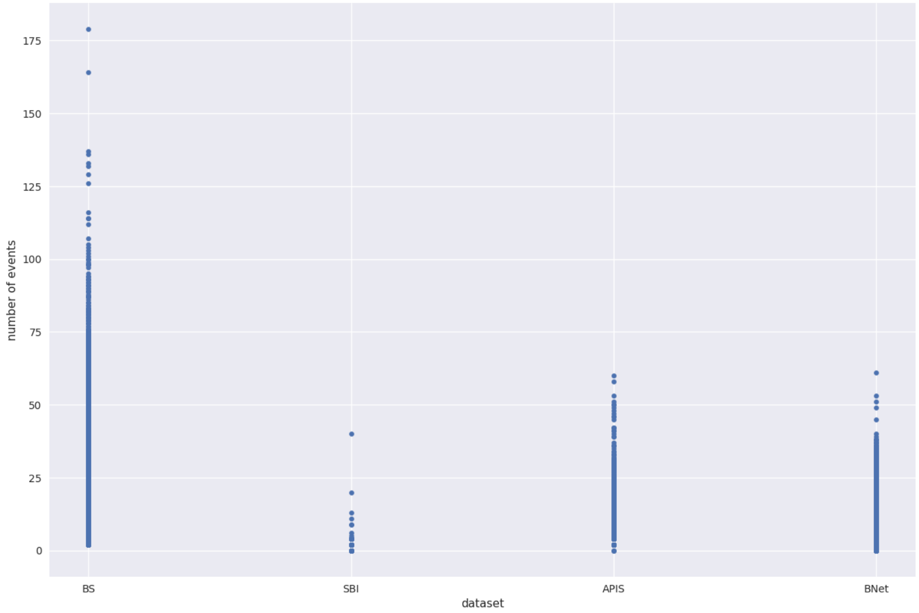

Fig. 3. Number of events per person in the four biographical data sources

Fig. 3 shows a scatter plot of the events per person. BS contains up to 175 events per person and includes a significant number of persons with more than 50 events. BNet and APIS on the other hand have only max 60 events per person and a significant number of people with around 25. The majority of data in SBI includes only death and birth events so far (with some manually enriched outliers). It has to be noted that Fig. 3 only includes the events that have been imported with the source datasets. The enrichment via the CHO pipeline components have added a massive amount of creation events to the IKG. The data for Gustav Klimt for example contains after the enrichment 140 events he participated as *producing artist* in.

One of the findings we didn't expect was the limited number – apart from APIS and SBI – of *sameAs* links between the datasets after ingestion (see Table 1). To improve the links between datasets we created Prefect enrichment pipelines using Wikidata to pull in more *sameAs* links (via additional identifiers) and events to create a denser dataset. These improvements are not yet reflected in Table 1.

Table 1. Number of *sameAs* links between the datasets

	APIS	**BS**	**BNet**	**SBI**
APIS	-	2	14	516
BS	2	-	9	0
BNet	14	9	-	0
SBI	516	0	0	-

9 Discussion and outlook

The InTaVia project shows that while technically it is possible to serialize datasets in a common format and merge them to a useful knowledge graph, many issues remain. Especially the automatic linking of vocabularies/concepts across languages and time is still an unresolved problem.

By evaluating the current status of the IKG we also found that the data quality needs further improvement. The source datasets are still not completely equally serialized – that is why we decided to implement the SHACL validation step. Some datasets contain not much structured data and/or the structured data could be improved and the IKG is not as dense as we expected it to be. However, the project also came up with some innovative approaches that we believe are of interest to other DH projects.

The combination of CIDOC CRM, Bio CRM and the ideas of linking *sameAs* entities via Provided_Entity instances taken from Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[39] allows for very flexible querying of the data via SPARQL and/or API. Users can select datasets they trust in and limit the statements on an entity to those datasets.

The infrastructure, a mixture of self-hosted ETL pipelines and GitHub actions proved very useful. It combines the benefits of publicly available code and data with the advantages of being able to execute long running, computing intense jobs on self-hosted infrastructure. Issues, code and commits are publicly available and therefore the data processing is reproducible, while the jobs themselves can be executed in on-premise infrastructure.

Finally, the implementation of the Rest API is based on the extension of FastAPI and the Pydantic classes allows for easy inline documentation of the data, simpler frontend development (due to the API contract via OpenAPI 3) and also simpler re-use of the data for other 3rd party applications and/or researchers not trained in SPARQL.

References

1. The Museo del Prado's knowledge graph. Retrieved 15-11-2023 (2023), https://www.museodelprado.es/en/grafo-de-conocimiento/el-grafo-de-conocimiento-del-museo-del-prado

[39] https://www.openarchives.org/ore

2. de Boer, V., van Rossum, M., Leinenga, J., Hoekstra, R.: Dutch ships and sailors linked data. In: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13. pp. 229–244. Springer (2014)
3. de Boer, V., Wielemaker, J., van Gent, J., Oosterbroek, M., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Amsterdam museum linked open data. Semantic Web **4**(3), 237–243 (2013)
4. Dijkshoorn, C., Aroyo, L., van Ossenbruggen, J., Schreiber, G.: Modeling cultural heritage data for online publication. Applied Ontology **13**(4), 255–271 (2018)
5. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75–92 (2003)
6. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The Europeana data model (EDM). In: World Library and Information Congress: 76th IFLA general conference and assembly. vol. 10, p. 15 (2010)
7. Erjavec, T., Dokler, J., Ogrin, P.V.: Slovenian biography. In: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017). pp. 16–21. CEUR Workshop Proceedings (2018), https://ceur-ws.org/Vol-2119/paper3.pdf
8. Fokkens, A., Ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V.: BiographyNet: Extracting relations between people and events. arXiv preprint arXiv:1801.07073 (2018)
9. Haslhofer, B., Isaac, A.: data.europeana.eu: The Europeana linked open data pilot. In: International conference on Dublin Core and metadata applications. pp. 94–104 (2011)
10. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., et al.: CultureSampo—Finnish culture on the semantic web 2.0. thematic perspectives for the end-user. In: Proceedings, museums and the web. pp. 15–18 (2009)
11. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo - publishing and enriching biographies on the semantic web for digital humanities research. In: The Semantic Web. ESWC 2019. pp. 574–589. Springer-Verlag (June 2019). https://doi.org/10.1007/978-3-030-21348-0_37
12. Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., Hyvönen, E.: WarSampo knowledge graph: Finland in the second world war as linked open data. Semantic Web **12**(2), 265–278 (2021). https://doi.org/10.3233/SW-200392
13. Lagoze, C., Van de Sompel, H., Nelson, M.L., Warner, S., Sanderson, R., Johnston, P.: Object re-use & exchange: A resource-centric approach. arXiv preprint arXiv:0804.2273 (2008)
14. Leskinen, P., Hyvönen, E.: Reconciling and using historical person registers as linked open data in the AcademySampo portal and data service. In: International Semantic Web Conference. pp. 714–730. Springer (2021)
15. Mäkelä, E., Hypén, K., Hyvönen, E.: BookSampo—lessons learned in creating a semantic portal for fiction literature. In: International Semantic Web Conference. pp. 173–188. Springer (2011)
16. Ockeloen, N., Fokkens, A., Ter Braake, S., Vossen, P., de Boer, V., Schreiber, G., Legêne, S.: BiographyNet: Managing provenance at multiple levels and from different perspectives. In: LISC@ ISWC. pp. 59–71 (2013)
17. Schlögl, M., Lejtovicz, K.: Die APIS-(Web-)Applikation, das Datenmodell und System. In: The Austrian Prosopographical Information System (APIS): vom gedruckten Textkorpus zur Webapplikation für die Forschung, pp. 31–48. NAP, New Aca-

demic Press (2020), https://www.oeaw.ac.at/fileadmin/Institute/ACDH/OEBL/pdf/_apis_Buch_WEB.pdf

18. Schlögl, M., Windhager, F., Mayr, E., Kaiser, M.: Biographische Informationssysteme (DPBs, Digital Knowledge Databases, Virtual Research Environments) (Mar 2019). https://doi.org/10.5281/zenodo.2593761
19. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the Smithsonian American Art Museum to the linked data cloud. In: The Semantic Web: Semantics and Big Data. pp. 593–607. Springer Berlin Heidelberg (2013). https://doi.org/10.1007/978-3-642-38288-8_40
20. Tuominen, J., Hyvönen, E., Leskinen, P.: Bio CRM: A data model for representing biographical data for prosopographical research. In: Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017). pp. 59–66. CEUR Workshop Proceedings (2018), https://ceur-ws.org/Vol-2119/paper10.pdf