

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

Matthias Schlögl¹[0000–0003–1451–0987], Jouni Tuominen^{2,3}[0000–0003–4789–5676], Joonas Kesäniemi³[0000–0002–3770–0006], Petri Leskinen^{2,3}[0000–0003–2327–6942], Go Sugimoto⁴[0000–0003–2646–6784], and Victor de Boer⁴[0000–0001–9079–039.X]

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

Abstract. We present the InTaVia Knowledge Graph (IKG), a large Knowledge Graph containing heterogeneous multilingual data from four European national biographies, connected to related cultural heritage objects. The InTaVia Knowledge Graph provides researchers, heritage professionals, and the informed public access to such biographical information. We describe the source data, the data model, and the resulting knowledge graph, as well as the pipeline components for managing and harmonizing the data. The data model combines domain standards CIDOC CRM and Bio CRM with elements to represent multiple perspectives on biographical information. The knowledge graph was consolidated from four prosopographical databases (PDBs) and enriched with links to Cultural Heritage Objects (CHOs) from Europeana and Wikidata. The resulting knowledge graph as information about 112,050 persons, described by 257,673 person proxies. In addition to the data model and the data itself, we also describe the infrastructure used to harmonize and maintain this heterogeneous knowledge graph.

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

Resource type: Dataset

License: CC BY 4.0 International

DOI: 10.5281/zenodo.10534980

URL: <https://qllever-ui.acdh-ch-dev.oeaw.ac.at/intavia>

1 Introduction

Biographical data consists of structured or unstructured information about the lives of individuals, including details such as birth and death dates, family relationships, educational and professional trajectories, achievements, and social

networks. This type of data is often sourced from historical records, literary texts, archival documents, and modern databases, and several of these biographical databases have been made available online.

An integrated knowledge graph of such biographical datasets offers a unified framework for representing and interlinking complex historical and cultural narratives, enabling researchers in digital humanities to uncover previously hidden patterns and relationships across diverse sources. By adhering to semantic web standards, such as RDF and OWL, these knowledge graphs facilitate interoperability and machine-readable data integration, making them a cornerstone for advancing linked data initiatives.

The main aims of the European project "In/Tangible European Heritage - Visual Analysis, Curation, and Communication" (InTaVia)⁵ are integrating structured data from four national biographical dictionaries, enriching these data with cultural heritage objects (CHOs) from reference resources, and providing a web-based visual analytics component that allows one to gain new insights in the data. InTaVia brings together data from four national biographical dictionaries: Austria (APIS), Finland (BiographySampo), Slovenia (SBI) and the Netherlands (BiographyNet). Since 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 intact. 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. For the digital humanities (DH), this integration supports enhanced data visualization, textual analysis, and computational modeling, providing new methodologies to explore and interpret biographical data at scale. In addition to the data itself, we argue that the data model and the infrastructure used to harmonize the data are of interest, especially to the DH community.

In the semantic web field, biographical knowledge graphs serve as rich, domain-specific exemplars of how structured data can improve information retrieval, natural language understanding, and cross-disciplinary knowledge discovery. It provides a central and open resource to connect other biographical or other historical data to a variety of applications and can serve as a benchmark dataset for (machine learning) methods and tools that deal with such data.

This paper covers the IKG as well as the conversion and data harmonization infrastructure used to construct and maintain it. We 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 that contain all the source code, most of the data and the issues that were used to discuss and decide on problems are available

⁵ <https://intavia.eu>

via our GitHub organisation⁶. Other available resources include the frontend application that allows to query and visualize the data (see Section 6)⁷, the API that is consumed by the frontend, but can also be used by other applications/users⁸ and the read-only SPARQL endpoint⁹.

In the next section, we position the IKG in the context of related work. Section 3 describes the IDM-RDF ontology used to harmonize the heterogeneous datasets. Section 4 then presents the four main parts of the IKG as well as general statistics. Section 5 describes the infrastructure, including data ingestion, conversion, and enrichment pipeline. Section 6 describes several use cases, and we present discussion and conclusions in Section 7.

2 Related work

Several Knowledge Graphs have been published in the domain of cultural heritage and digital humanities that present rich information from a specific heritage institute. Examples include the Rijksmuseum Linked Data [5], the Prado Knowledge Graph [19], the Smithsonian American Art Museum Linked Data [23], 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 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 (mainly Finnish) cultural heritage and history using knowledge graphs. These aspects range from culture [11], literature [18], war history [13] or academic history [17]. Several of these knowledge graphs explicitly model biographical data. The Dutch Ships and Sailors graph [1] does this for several Dutch maritime historical datasets. Europeana provides access to its aggregated collections through a SPARQL endpoint [10]. 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 intact 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 [21] was used. APIS is based on a relational database, but allows us 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).

⁶ <https://github.com/intavia>

⁷ <https://intavia.acdh-dev.oeaw.ac.at>

⁸ <https://intavia-backend.acdh-dev.oeaw.ac.at/v2/docs>

⁹ <https://qllever-ui.acdh-ch-dev.oeaw.ac.at/intavia>

In general, for publishing biographical/prosopographical data a variety of technologies have been used: TEI (such as the Slovenska biografija [8]), relational databases (such as the APIS dataset), document-based databases/search indexes (such as the Neue Deutsche Biographie¹⁰) and RDF-based systems (such as BiographyNet [9] and BiographySampo [12]) (see [22] 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 [20] the ability to describe multiple perspectives on persons, via the ORE-OAI Proxy model [15], which itself was borrowed from the Europeana Data Model [7]. However, the main part of our data model is based on the domain standard CIDOC CRM [6]. Its event-centric model is very suitable for describing a variety of heterogeneous objects. The Bio CRM extension provides additional constructs and design patterns to describe biographical information, specifically towards prosopographical descriptions [24].

3 The ontology (IDM-RDF)

In this section, we describe the ontology used to harmonize the various datasets. The InTaVia Data Model in RDF (IDM-RDF) was developed iteratively in collaboration between the project partners to meet the requirements of the source prosopographical databases (PDBs) of the project partners, the cultural heritage object databases (ODB) and potential users. The data model allows us to express basic prosopographical and object data from one or more potentially conflicting sources with annotated source texts and provenance data for the bibliographic sources and for the automated processing steps of various included data sets. The IDM-RDF is based on a modular approach. In the following sections, we describe the main components.

Figure 1 shows an overview of how actors are related to cultural heritage objects in IDM-RDF. The complete IDM-RDF data model is published on GitHub¹¹, with archival on Zenodo¹².

3.1 CIDOC CRM as core

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 center the IDM-RDF ontology around the CIDOC CRM v7.1.1 implemented in RDFS¹³. The main reasons for this is that CIDOC CRM is a recognized (ISO) standard for the domain, is under

¹⁰ http://www.ndb.badw-muenchen.de/ndb.aufgaben_e.htm

¹¹ <https://github.com/InTaVia/idm-rdf>

¹² <https://doi.org/10.5281/zenodo.5534542>

¹³ https://cidoc-crm.org/rdfs/7.1.1/CIDOC_CRM_v7.1.1.rdfs



Fig. 1: Schema detail of the IDM-RDF that shows the relation of an actor and a cultural heritage object.

active development since the 1990s and widely adopted in the DH community. Furthermore, CIDOC CRM is already in use by two out of four biographical data providers (ACDH-CH, Aalto University).

3.2 Bio CRM

The extension Bio CRM [24] is used to model prosopographical data. It specifically is designed for biographical data and is compatible with the CIDOC CRM core. Bio CRM allows for modeling of prosopographical facts such as nationality, gender, social relations, and occupations. 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.

3.3 Provenance modeling

On a high level, platform provenance describes the processes of creating datasets and their transformations. In order to keep track of these processes and their related metadata, InTaVia utilizes PROV-O provenance ontology [16] and its P-PLAN extension¹⁴, which provides a concrete implementation for *prov:Plan* and allows to distinguish between planned and actualized executions.

¹⁴ <http://purl.org/net/p-plan#>

In the InTaVia bibliographic provenance graph, BIBFRAME¹⁵ is used to model the data about the source where the biographic information comes from originally, up to the level of an article. A relation between the *idm:Person_Proxy* (which represents the perspective of this source on a person’s biography) and the biography article makes this data available for querying. BIBFRAME was developed by the Library of Congress since 2011 and is updated regularly.

3.4 Multiple perspectives using OAI-ORE proxies

As we have information about persons, groups and cultural heritage objects from various sources, IDM-RDF needs to support integrating potentially contradictory information, but also ensure that these various perspectives remain accessible. To this end, we reuse the proxy concept from Open Archives Initiative Object Reuse and Exchange (OAI-ORE)¹⁶. OAI-ORE “defines standards for the description and exchange of aggregations of Web resources”. OAI-ORE’s proxy concept was earlier adapted and modified, for example, by Europeana [7] and BiographyNet [20]. In IDM-RDF, its application is even more simplified. In the OAI-ORE definition the necessity of a *proxy* is defined for the case that “this fact is only true in the context of the specific Aggregation, and is not a ‘global’ fact”. In the context of humanities, “global facts” are rare. The diversity and inconsistency of data that was collected over a large time span can be relevant for research and therefore it is important to have access to that data and its provenance data. In InTaVia, an `ore:Proxy` stands for a certain perspective on a person, group or cultural heritage object in the context of a specific source.

3.5 Europeana Data Model

The modeling of the CHO data was done after detailed consideration of the Europeana Data Model (EDM) [7] and exemplary queries of the Europeana data via the official Europeana SPARQL endpoint¹⁷.

4 The InTaVia Knowledge Graph

The IKG can be queried and retrieved via the REST API or SPARQL endpoint listed in Section 1. Additionally, current versions of prosopographical datasets are stored in the source data repository¹⁸. This data repository is hooked to Zenodo and archived on every new release¹⁹. Smaller datasets, e.g. CHO data and person relations enrichment data from Wikidata, is stored in the source-dataset-conversion²⁰ repository. The IKG currently contains close to 22 000 000 triples

¹⁵ <https://www.loc.gov/bibframe/docs/index.html>

¹⁶ <https://www.openarchives.org/ore/>

¹⁷ <http://sparql.europeana.eu>

¹⁸ <https://github.com/InTaVia/source-data>

¹⁹ <https://zenodo.org/doi/10.5281/zenodo.10290205>

²⁰ <https://github.com/InTaVia/source-dataset-conversion>

which describe roughly 260 000 proxies. We include a VOID file describing the dataset (https://github.com/InTaVia/source-data/blob/main/void_graph_intavia.ttl).

4.1 The four biographical dictionaries in IKG

InTaVia brings together data from four national biographical dictionaries. Here we describe these four core parts of the IKG: APIS, BiographySampo, Slovenia and BiographyNet.

APIS The APIS dataset created from the Austrian Biographic Dictionary (ÖBL)²¹ 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 significance of the source dataset.

Of the 18 332 places entities of the source data set 7019 are included in the InTaVia triplestore, those with relationships to persons and / or institutions BL. 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 that contain relations to institutions.

The APIS data includes data created by researchers through manual annotations. While the original API provides provenance data on who created these annotations, the IKG serialization currently misses these provenance metadata.

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 Society²², 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 the names of persons are modeled with 33 944 resources in the class *E33 E41 Linguistic Appellation*. Furthermore, the data contains approx. 103 000 lifetime events of the actors modeled 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

²¹ <https://www.oeaw.ac.at/acdh/oebl>

²² <https://kansallisbiografia.fi/english>

represented by 3050 entities of the class *E36 Visual Item*. Altogether 5642 people have a link (*owl:sameAs*) to a corresponding Wikidata entity. Furthermore, the IKG currently contains 797 BiographySampo occupation categorizations for persons and 2745 family relationship roles.

BiographyNet The IKG contains all 79 412 person entities from BiographyNet (BNet). This dataset is the only main source dataset that contains potentially multiple biographical descriptions of the same person. While the data had previously been converted to Linked Data, for the harmonization, the original source data from the Biography Portal of the Netherlands (BPN)²³, 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, as defined by the proxy construction 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 are still modeled with classes from the specific BNet data model.

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, and motion pictures. Events include memberships and work-related postings.

4.2 IKG statistics

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

²³ <http://www.biografischportaal.nl/en/>

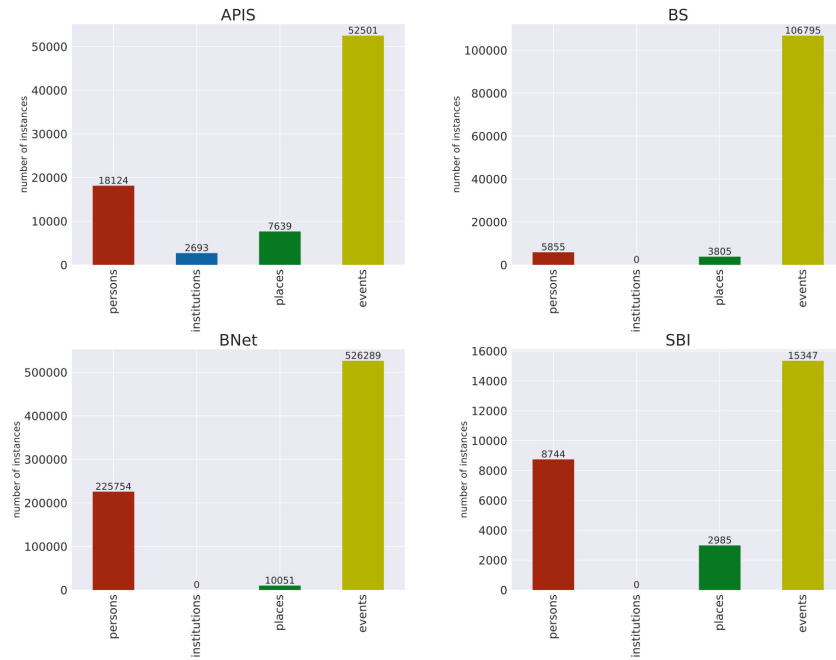


Fig. 2: Number of entities in the four biographical data sources

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.

One of the findings we did not expect was the limited number of *sameAs* links between the datasets after ingestion. Only 548 of such *sameAs* links were found between persons from different datasets, with the majority (517) being between SBI and APIS. To further connect the graph, we created Prefect enrichment pipelines using Wikidata to pull in more *sameAs* links (via additional identifiers) and events to create a denser dataset. In total, the IKG contains 328 693 *sameAs* links, that connect entities both internally and to external sources.

5 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 Blazegraph²⁴ triplestore in quad mode as the database backend of the API.
- A FastAPI²⁵ 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 Prefect²⁶ 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)
- Qlever triplestore with Qlever-UI²⁷ frontend for direct access to the data via SPARQL

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 Kubernetes²⁸ cluster.

5.1 ETL pipeline

For serializing the datasets, and for further enrichment and curation tasks, an ETL framework was developed based on Prefect²⁹. To ensure sustainability of the resource and to allow for constant updating of the IKG in case the source data is updated, an additional workflow has been implemented in GitHub. Using GIT LFS and GitHub workflows for SHACL validation³⁰, this approach allows to not only keep a record of the previous states of the knowledge graph, but also discuss possible problems of certain datasets in PRs.

The source-data repository³¹ is configured to allow for the following workflow. 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 kept in the 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. The main branch is used to populate the production triplestore with new data.

²⁴ <https://blazegraph.com>

²⁵ <https://fastapi.tiangolo.com>

²⁶ <https://www.prefect.io>

²⁷ <https://github.com/ad-freiburg/qlever>

²⁸ <https://kubernetes.io>

²⁹ Prefect is a Python-based ETL framework. <https://www.prefect.io>

³⁰ <https://www.w3.org/TR/shacl/>

³¹ <https://github.com/InTaVia/source-data>

This approach offers a central tool for data validation, issue tracking and versioning of the data while limiting the infrastructure requirements.

Workflows for creating datasets, enriching the IKG, or creating the links between the entities have been implemented with Prefect³². In the following we will discuss the core workflows.

5.2 Dataset conversions

Each of the four source datasets was converted using a dataset-specific strategy. For example, BNet was converted from its original XML format to IDM-RDF through a Prolog-based pipeline based on tooling described in [2]. The BiographySampo data was already available in RDF format and needed only slight adaptations. APIS was converted to IDM-RDF using a Prefect workflow. The workflow³³ pulled data from the REST API APIS provides and converted it to IDM-RDF. SBI was converted using a standalone Python script³⁴ from TEI to IDM-RDF. The SBI conversion has not been ported to prefect as the original TEI files are not available. For all of the 4 biographical datasets the GitHub based validation process described above was used to iteratively find and resolve inconsistencies.

5.3 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 sameAs 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 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.

5.4 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,

³² The workflows used for creating the stable IKG are based on Prefect and are published in <https://github.com/InTaVia/prefect-flows>. Some of the workflows have been upgraded to v2 and published in <https://github.com/InTaVia/prefect2-flows>

³³ https://github.com/InTaVia/prefect2-flows/blob/main/create_apis_graph_v3.py

³⁴ https://github.com/InTaVia/source-dataset-conversion/tree/main/SBI_dataset

only the relations between IKG actors were chosen. The interperson relations were related to 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 was added to its own named graph in the InTaVia triplestore. Similarly, the interperson relations extracted from Getty Union List of Artists' Names have been added to its own named graph.

5.5 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. Not to run into timeouts, it uses a configurable while loop to iterate through all persons with Wikidata identifier in the IKG. The enrichment flow queries for any objects connected to the person via *wdt:P170* (creator). It uses the object titles, 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.

5.6 Enrich the IKG with cultural heritage objects from Europeana

In the project plan, Europeana was foreseen as the main source of data on cultural heritage. 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 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.

6 Use cases

6.1 The InTaVia frontend and API

The main use case for the IKG is the InTaVia platform, which provides a frontend for accessing the data. The platform was developed to support data-driven storytelling through data retrieval, creation, curation, analysis, and communication with coherent visualization support for multiple types of entities [14]. The frontend, which is accessible at <https://intavia.acdh-dev.oeaw.ac.at/> supports three types of uses: 1) Search & Curation, 2) Visualization & Analysis and 3) Storytelling & Presentation for various types of users. The target user group is varied and includes (Digital) Humanities scholars, educators, and journalists.

Fig. 3 shows two example screenshots of the InTaVia frontend. The screenshot on the left shows the search and curation interface, displaying results for

the person Giuseppe Acerbi. It shows the timeline of life events, sourced from multiple sources as well as geographical relations³⁵. The right image shows an example of the storytelling suite, where a presentation can be found of the life and times of Pier Paolo Vergerio³⁶



Fig. 3: Two example screenshots of the InTaVia platform

The frontend runs on the IKG, with a bespoke REST API on top of the SPARQL endpoint. Such an API middle-layer allows for performance optimizations and ease of (re)use for application development. The API is documented at <https://intavia-backend.acdh-dev.oeaw.ac.at/v2/docs> using OpenAPI 3 specification of parameters³⁷. 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 API delivers results from the IKG in a JSON format that includes 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.

The API is built using FastAPI, a Python framework for building REST APIs. The core functionality of the API is a pydantic model that allows to build arbitrary deeply nested JSON objects out of the flat SPARQL JSON directly from within the pydantic model definition³⁸.

³⁵ <https://tinyurl.com/InTaViaExample1>

³⁶ <https://tinyurl.com/InTaViaExample2>

³⁷ The source code is available in <https://github.com/InTaVia/InTaVia-Backend>

³⁸ see <https://github.com/InTaVia/InTaVia-Backend> for implementation details.

6.2 Reconciliation API

Although the IKG is available for re-use outside of the project context through both the public SPARQL endpoint as well as the JSON API discussed in the previous section, this information is also available for entity reconciliation services via a standardized API, allowing for easier re-use in tools and applications other than those of InTaVia. The Reconciliation Service API [4] specifies a protocol for data matching on the Web, especially tailored for the Semantic Web technologies. The API specifications are drafted as a W3C community group report. The API allows for batch lookup of entities for the purpose of entity reconciliation. Tools such as OpenRefine support these services. It has previously been implemented by entity data providers such as Wikidata, RKD Artists, Getty Vocabularies, and VIAF among others. The code of the InTaVia Reconciliation



Fig. 4: Example use of the IKG data in OpenRefine through the reconciliation API

Service API is a part of InTaVia Backend³⁹. The API itself is available at <https://intavia-backend.acdh-dev.oeaw.ac.at/v1/recon/reconcile/>. Queries are sent via HTTP POST method with information about a query term (e.g. “Franz”), a maximum number of results (e.g. 10), and an entity type (e.g. “Person”). The response is a JSON list of results containing the matching entity URI, a matching score (0–1) and the main label of the entity (for persons, the first and last name). Fig. 4 shows how the this reconciliation API is used in OpenRefine to reconcile a person name.

6.3 Example SPARQL queries

Of course, the raw data in the knowledge graph can be accessed through the aforementioned SPARQL endpoints. We here list a few representative SPARQL queries that can be used on the data to retrieve relevant datasets, answer digital humanities questions or show statistics over the dataset.

³⁹ <https://github.com/InTaVia/InTaVia-Backend>

- <https://qllever-ui.acdh-ch-dev.oeaw.ac.at/intavia/mmFCS2> runs a query that searches in the biographical datasets for a person named "Acerbi" and returns the URIs, labels, and datasets of the found persons. The query can be easily adapted to other names (or all overlaps by removing the filter).
- <https://qllever-ui.acdh-ch-dev.oeaw.ac.at/intavia/FTqCWc> shows a query for *owl:sameAs* links from Persons in the datasets. These *sameAs* links have been used to link the entities from the various datasets together.
- <https://qllever-ui.acdh-ch-dev.oeaw.ac.at/intavia/1cg6K> counts all entities in the knowledge graph. It has to be noted that these entities are proxies in the KG. As there can be multiple proxies for a single person, place or group, this number is larger than the number of distinct entities in the IKG.
- <https://qllever-ui.acdh-ch-dev.oeaw.ac.at/intavia/Fh9cNW> retrieves the number of people (born after 1800) for each registered profession from the BiographyNet graph only. This query was constructed at the request of a Dutch DH scholar interested in historical occupations.

7 Discussion and conclusions

The IKG is a rich knowledge graph that integrates heterogeneous prosopographical and cultural heritage object data from various European sources. This presents a valuable resource for Digital Humanities use cases. The combination of CIDOC CRM, Bio CRM and the ideas of linking *sameAs* entities via *Provided_Entity* instances as defined by OAI-ORE allows for very flexible querying of the data via SPARQL and/or API. Users can select datasets they trust and limit the statements on an entity to those datasets. We also presented use cases via a web frontend and a reconciliation API.

While we show that it is possible to serialize datasets in a common format and merge them to a useful knowledge graph, several opportunities for further enrichment are identified. 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. However, the project also came up with some innovative approaches that we believe are of interest to other DH projects.

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, source code, and commits are publicly available and therefore the data processing is reproducible, while the jobs themselves can be executed on an on-premise infrastructure.

Acknowledgements The authors would like to thank Joh Dokler, Carla Ebel and the rest of the InTaVia consortium for the valuable collaboration. This work was funded by the EU H2020 research and innovation action InTaVia, project No. 101004825.

References

1. 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)
2. de Boer, V., Wielemaker, J., van Gent, J., Hildebrand, M., Isaac, A., van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The Amsterdam Museum case study. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) *The Semantic Web: Research and Applications*. pp. 733–747. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
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. Delpuch, A., Pohl, A., Steeg, F., Guidry Sr., T., Suominen, O.: Reconciliation Service API v0.2: A protocol for data matching on the web. Final Community Group Report 10 April 2023, W3C (2023), <https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/>
5. Dijkshoorn, C., Aroyo, L., van Ossenbruggen, J., Schreiber, G.: Modeling cultural heritage data for online publication. *Applied Ontology* 13(4), 255–271 (2018)
6. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine* 24(3), 75–92 (2003)
7. 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)
8. 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>
9. Fokkens, A., Ter Braake, S., Ockeloën, N., Vossen, P., Legêne, S., Schreiber, G., de Boer, V.: BiographyNet: Extracting relations between people and events. arXiv preprint arXiv:1801.07073 (2018)
10. 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)
11. 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)
12. 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
13. 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>
14. Kusnick, J., Mayr, E., Seirafi, K., Beck, S., Liem, J., Windhager, F.: Every thing can be a hero! narrative visualization of person, object, and other biographies. *Informatics* 11(2) (2024). <https://doi.org/10.3390/informatics11020026>, <https://www.mdpi.com/2227-9709/11/2/26>

15. 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)
16. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013 (2013), <https://www.w3.org/TR/2013/REC-prov-o-20130430/>
17. 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)
18. 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)
19. The Museo del Prado’s knowledge graph. Retrieved 15-11-2023 (2023), <https://www.museodelprado.es/en/grafico-de-conocimiento/el-grafico-de-conocimiento-del-museo-del-prado>
20. Ockeloen, N., Fokkens, A., Ter Braake, S., Vossen, P., de Boer, V., Schreiber, G., Legène, S.: BiographyNet: Managing provenance at multiple levels and from different perspectives. In: LISC@ ISWC. pp. 59–71 (2013)
21. 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 Academic Press (2020), https://www.oeaw.ac.at/fileadmin/Institute/ACDH/OEBL/pdf/_apis_Buch_WEB.pdf
22. 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>
23. 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
24. 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>