Journal Title
XX(X):1–10

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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

Matthias Schlögl[1], Jouni Tuominen[2,3], Joonas Kesäniemi[3], Petri Leskinen[2,3], Go Sugimoto[4] and Victor de Boer[4]

Abstract

The InTaVia Knowledge Graph (IKG) is a large Knowledge Graph containing heterogeneous multilingual data from four European national biographies, connected to related cultural heritage objects. This resource provides researchers, heritage professionals, and the informed public access to such biographical information. This paper describes the source data, the data model, the pipeline components for managing and harmonizing the data and the resulting knowledge graph. 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

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 dictionaries offers a unified framework for representing and interlinking complex historical and cultural narratives, enabling researchers in the 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. Furthermore, we can use these ontology standards to specify formal semantics of biographical domain relations, allowing for automated reasoning.

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 were to be integrated are highly heterogeneous. We therefore turned to semantic web technology to bring these datasets together into a single integrated knowledge graph while keeping their richness intact. This paper presents the resulting 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. 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.

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

[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]https://intavia.eu

data and the issues that were used to discuss and decide on problems are available via our GitHub organisation[6]. Other available resources include the frontend application that allows to query and visualize the data (see the Use cases Section)[7], the API that is consumed by the frontend but can also be used by other applications/users[8] and the read-only SPARQL endpoint[9].

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 [1], the Prado Knowledge Graph [2], the Smithsonian American Art Museum Linked Data [3], or the Amsterdam Museum knowledge graphs [4]. 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 above-mentioned 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 [5], literature [6], war history [7] or academic history [8]. Several of these knowledge graphs explicitly model biographical data. The Dutch Ships and Sailors graph [9] 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 [11] 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).

In general, for publishing biographical/prosopographical data a variety of technologies have been used: TEI[12] (such as the Slovenska biografija [13]), relational databases (such as the APIS dataset), document-based databases/search indexes (such as the Neue Deutsche Biographie[10]) and RDF-based systems (such as BiographyNet [14] and BiographySampo [15]) (see [16] 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 [17] the ability to describe multiple perspectives on persons, via the ORE-OAI Proxy model [18], which itself was borrowed from the Europeana Data Model [19]. However, the main part of our data model is based on the domain standard CIDOC CRM [20]. 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 [21].

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.

These requirements include that the data model must be capable of expressing all relevant information present in the source datasets, provide the data in formats that fit the needs of and follow general best-practices and standards and 1) Cover and combine cultural heritage object data and biographical data; 2) Combining different data sources; 3) Implementation of vocabularies and modularized IDM-RDF approach, and 4) The model has to provide a comprehensible representation of contradictory data with provenance of information, both for the original sources and for data processing steps (NLP). Incomplete and uncertain data must be modeled in a transparent and comprehensible manner to enable visualization of uncertainties."

Based on these requirements, we developed a modular data model consisting of several (re-used) components, which are described below. Figure 1 shows an overview of the datamodel and how an actor is related to several life events as well as to cultural heritage objects in IDM-RDF. The complete IDM-RDF data model is published on GitHub[11], with archival on Zenodo[12].

CIDOC CRM as core

We center the IDM-RDF ontology around the CIDOC CRM v7.1.1 implemented in RDFS[13]. The main reasons for this is that CIDOC CRM is a recognized (ISO) standard for the domain, is under 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).

[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://qlever-ui.acdh-dev.oeaw.ac.at/intavia
[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

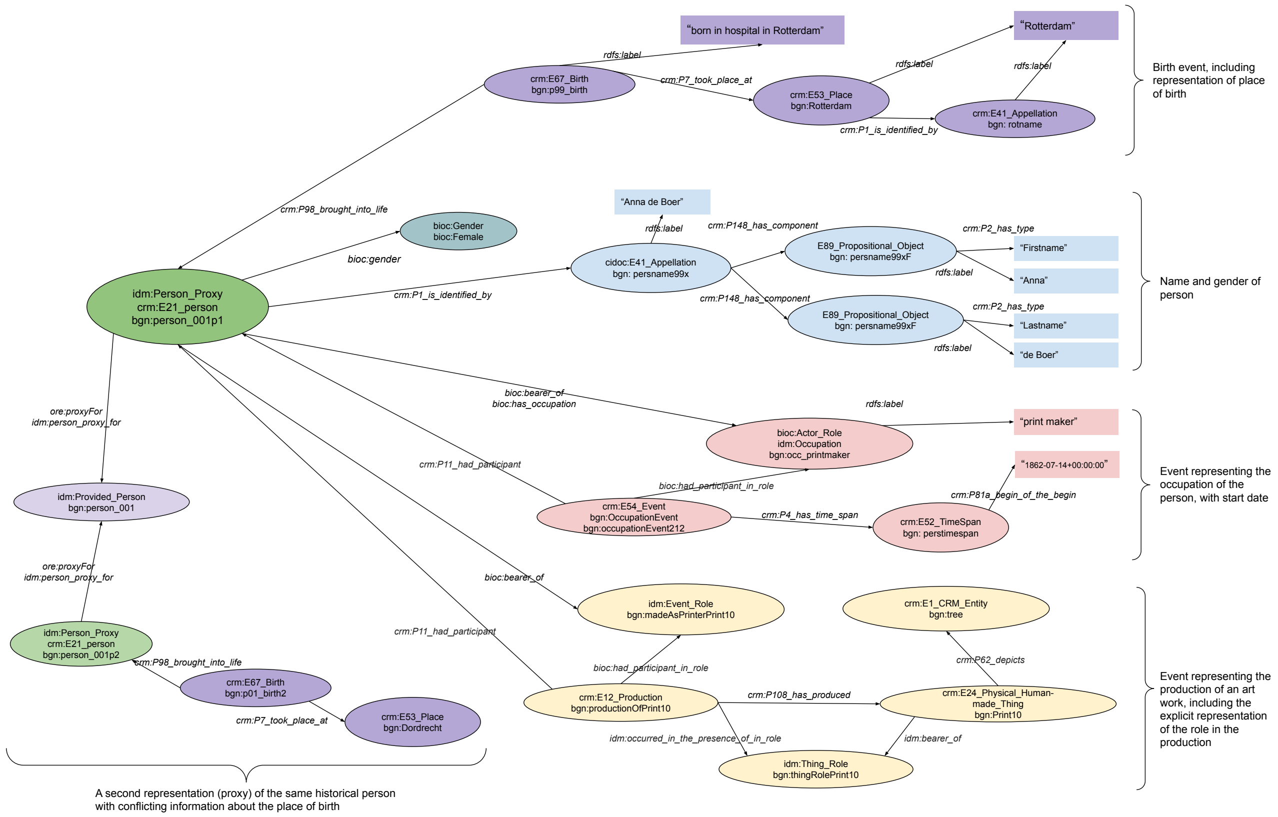


Figure 1. Visualisation of a modelled example in IDM. Ellipses represent resources, with class URIs written above the instance URI. The image shows a fake example of a person named Anna de Boer, a birth event and an occupation event as well as the (creation) relation between the person of an actor and a cultural heritage object. The namespaces used in the figure are crm: http://www.cidoc-crm.org/cidoc-crm/, idm: http://www.intavia.eu/idm-core/, bioc: http://ldf.fi/schema/bioc/ and bgn: http://data.biographynet.nl/rdf/

Bio CRM

The extension Bio CRM [21] is used to model prosopographical data. It is specifically 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 the roles of actors and things in events and to allow the implementation of 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.

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 the PROV-O provenance ontology [22] and its P-PLAN extension[14], which provides a concrete implementation for *prov:Plan* and allows one to distinguish between planned and actualized executions[15].

In the InTaVia bibliographic provenance graph, BIBFRAME[16] 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.

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)[17]. OAI-ORE "defines standards for the description

[14] http://purl.org/net/p-plan#

[15] PROV-O was selected over other provenance representation standards, including CRMdig because of its status as a W3C standard and its integration with the P-Plan ontology.

[16] https://www.loc.gov/bibframe/docs/index.html

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

and exchange of aggregations of Web resources". OAI-ORE's proxy concept was earlier adapted and modified, for example, by Europeana [19] and BiographyNet [17]. 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. For example, the BiographyNet dataset contains many cases where several biographies exist of the same person, and in some cases this leads to differences in the metadata (e.g. different birth places or name spellings). This can then be adequately modeled through different proxies linked to the same provided person (See Figure 1, bottom left). The other main use case is for InTaVia enrichments. Rather than amending the metadata for the existing proxy, we generate a new proxy to which we add the new enriched/generated metadata, thereby keeping the original data and new data separate but linked.

Europeana Data Model

The modeling of the CHO data was done after detailed consideration of the Europeana Data Model (EDM) [19] and exemplary queries of the Europeana data via the official Europeana SPARQL endpoint[18].

Automated reasoning

the IDM ontology and CIDOC CRM are expressed using OWL and RDFS. They use OWL constructs such as *owl:inverseOf*, *rdfs:subClassOf*, or *rdfs:subPropertyOf* to define the formal semantics of domain-specific properties and classes and support automated reasoning. For example: the property *crm:P17_was_motivated_by* is defined as a subPropertyOf *crm:P15_was_influenced_by*. So, if one uses an SPARQL query to retrieve people influenced by someone, we also find people motivated by them. In the InTaVia middle layer, typically, classes and properties higher in the hierarchy are used, such that the more specific classes and properties as they appear in the individually converted datasets are retrieved.

The InTaVia Knowledge Graph

The IKG can be queried and retrieved via the REST API or SPARQL endpoint listed in the Introduction section. Additionally, current versions of prosopographical datasets are stored in the source data repository[19]. This data repository is hooked to Zenodo and archived on every new release[20]. Smaller datasets, e.g. CHO data and person relations enrichment data from Wikidata, is stored in the source-dataset-conversion[21] repository. The IKG currently contains close to 22 000 000 triples which describe roughly 260 000 proxies. We include a VoID file describing the dataset[22]. The dataset is published under the Creative Commons Attribution 4.0 International license.

The four biographical dictionaries in IKG

InTaVia brings together data from four national biographical dictionaries: APIS, BiographySampo, Slovenia and BiographyNet. We describe them below.

APIS The APIS dataset created from the Austrian Biographic Dictionary (ÖBL)[23] 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 about 6512 actor entities covered in BiographySampo (BS), which is based on the National Biography of Finland and other biographical databases of the Finnish Literature Society[24], interlinked with related data repositories. People still alive, fictional characters, and actors representing families or kin are not included in the IKG. For all the included 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 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

[18] `http://sparql.europeana.eu`
[19] `https://github.com/InTaVia/source-data`
[20] `https://zenodo.org/doi/10.5281/zenodo.10290205`
[21] `https://github.com/InTaVia/source-dataset-conversion`
[22] `https://github.com/InTaVia/source-data/blob/main/void_graph_intavia.ttl`
[23] `https://www.oeaw.ac.at/acdh/oebl`
[24] `https://kansallisbiografia.fi/english`

	APIS	BS	Bnet	SBI
Persons	18124	5855	225754	8744
Institutions	2693	0	0	0
Places	7639	3805	10051	2985
Events	52501	106795	526289	15347
Birth Events	18106	5779	124399	8092
Birth Places	18093	5232	85067	8041
Birth Dates	18106	5774	105549	7949
Gender male	17201	5010	666	7902
Gender female	923	284	845	842
Persons with >0 occupation	18095	5498	26814	8145
Persons with at least >1 name component	18124	5820	20604	8664

Table 1. Number of entity proxies and additional statistics in the four biographical data sources

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)[25], 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 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. Additionally, the SBI subset of the data contains 4333 related person references and 119 family entries, which were not modeled as full biography proxies. 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.

IKG statistics

Table 1 gives an overview of the total number of entity proxies for each of the entity types in the source datasets. No duplicate detection or reconciliation is performed: if a person appears in two datasets they are counted twice. This corresponds to the principle of having entity proxies represent a perspective on a person rather than the person itself. The numbers show that the data are 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 type 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.

Table 1 furthermore shows that the completeness of the metadata varies per data set. For example, APIS is relatively complete, listing for over 95% of the Persons their birth events, gender, occupation, and first and last names. For BiographyNet this is less complete, where, for example, for only 55% of the person proxies a birth event is known. The statistics also show that most datasets are heavily biased towards male persons, with 85-95% of the persons explicitly listed as 'male'. For BiographyNet, most of the time, the gender is not known, but in cases where it is available, a slight majority is listed as 'female'.

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[26] 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.

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 1) a Blazegraph[27] triplestore in quad mode as the database backend of the API; 2) A FastAPI[28] backed REST API that consumes the triplestore and delivers JSON to the frontend (see also the Use Cases Section); 3) A React based frontend that allows to search the IKG and perform visual analytics

[25] http://www.biografischportaal.nl/en/
[26] Prefect is a Python-based ETL framework. https://www.prefect.io
[27] https://blazegraph.com
[28] https://fastapi.tiangolo.com

on the data; 4) A Prefect 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) 5) QLever triplestore with QLever UI[29] frontend for direct access to the data via SPARQL available at https://qlever-ui.acdh-dev.oeaw.ac.at/intavia. 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[30] cluster.

ETL pipeline

For serializing the datasets, and for further enrichment and curation tasks, an ETL framework was developed based on Prefect. 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. We use GIT LFS and GitHub workflows for SHACL validation[23]. This update pipeline in essence is a re-conversion of source data, rather than an update protocol. This means that as soon as changes are available, the ETL pipeline is run again, updating the IKG in its entirety. Therefore any changes to existing data are also overwritten, taking care of updated, deleted and new content. 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. For all of the 4 biographical we use SHACL and GitHub to validate the data matches the IDM ontology and to resolve inconsistencies.

Dataset conversions

Each of the four source datasets was converted using a dataset-specific strategy. Only for the generation of the BNet data a formal mapping language was used [24] to map the source XML data into IDM compliant RDF. For the generation of the APIS data the already existing JSON API was used and the RDF was built using RDFlib within a Prefect flow[31]. SBI was built also using RDFlib from the existing TEI files and BS needed only slight modifications (done with using a combination of RDFLib and SPARQLWrapper for accessing the BS data) as it was already modeled using CIDOC CRM.

In the three cases where the source data was not in IDM compliant RDF format, limited data cleaning and normalization was performed. This concerned the rewriting of the source TEI/JSON/XML into RDF triple format in the appropriate syntactic format: values are converted to RDF literals; relations and tags are mapped to either IDM properties or classes etc. Since the target data model mostly reuses CIDOC CRM, this conversion also meant introducing event-centric modelling for much of the metadata. For example, BiographyNet lists birthdates as simple relations between a person entry and a date value, whereas in the CIDOC model, this is represented using an explicit birth event resource (of type *E67_Birth*). The conversions all required such 'refactoring' of metadata in addition to more straightforward syntacting rewriting. For each dataset, these refactoring decisions are documented in their respective conversion scripts.

Inaccurate dates

We use CIDOC CRM's time-spans for inaccurate dates, using the two (not all the four) properties: *P82a_begin_of_the_begin* and *P82b_end_of_the_end*. E.g. if we know that some event has happened during the year 1983:P82a begin of the begin 1983-01-01T00:00:00Z and P82b end of the end 1983-12-31T23:59:59Z.

Entity reconciliation process

For reconciling entities (persons, places) in the four biographical source datasets, we have implemented an entity ID enrichment process. This process first takes into account existing sameAs relations that are found in the ingested data. For missing sameAs relations, we query the Wikidata SPARQL endpoint for additional 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.

Enrichments

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 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.

We also enrich the IKG with CHOs from Wikidata by using federated SPARQL queries against Wikidata to copy the needed data. 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.

Finally, data is enriched with Europeana CHOs. Due to technical challenges with the Europeana SPARQL endpoint, we opted to batch download the related object data via the Europeana API, converting to a turtle file and uploading it to the IKG. All enrichments are stored in separate Named Graphs to allow for easy management and provenance descriptions.

[29] https://github.com/ad-freiburg/qlever
[30] https://kubernetes.io
[31] https://github.com/InTaVia/prefect2-flows/blob/main/create_apis_graph_v3.py

Use cases

The InTaVia frontend and API

The main use case for the IKG is the InTaVia platform, which provides a frontend for accessing the data. Server logs suggest that we currently get between 300 and 600 requests per month. 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 [25]. The frontend[32] 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. 2 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 relations[33]. The right image shows an example of the storytelling suite, where a presentation can be found of the life and times of Pier Paolo Vergerio[34].

The frontend runs on the IKG, with a bespoke REST API[35] 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 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 definition[36].

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 [26] specifies a protocol for data matching on the Web, especially tailored to 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 provides such as Wikidata, RKD Artists, Getty Vocabularies, and VIAF among others. The code of the InTaVia Reconciliation Service API[37] is a part of InTaVia Backend[38]. 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. 3 shows how the this reconciliation API is used in OpenRefine to reconcile a person name.

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.

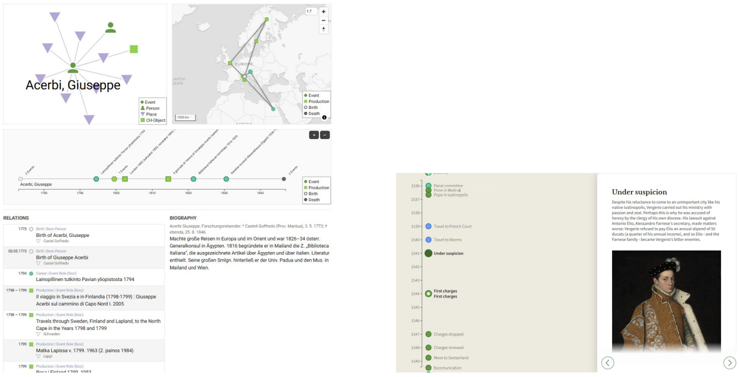
- https://qlever-ui.acdh-dev.oeaw.ac.at/intavia/qwQMka runs a query that searches in the biographical dictionaries for a person named "Acerbi" and returns the URIs, labels, and datasets of the found persons. The query can be easily adapted to other names (or all overlaps by removing the filter).
- https://qlever-ui.acdh-dev.oeaw.ac.at/intavia/q75psN 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://qlever-ui.acdh-dev.oeaw.ac.at/intavia/z5ZNoi 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://qlever-ui.acdh-dev.oeaw.ac.at/intavia/yQn3A2 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.

User Evaluations

To evaluate the InTaVia user platform, a workshop was held with 11 humanities experts [27]. While user tasks focused on the usability and transparency of the tools, we still report here the main outcomes, as the tools expose the content with the IKG. In the beginning, we asked for the participant's informed consent to document and analyze the workshop results. In the first workshop part, participants worked with google colab to modify and run prepared requests to the InTaVia backend and visualize the results. In the second part of the workshop, we moved on to the InTaVia platform, which was introduced step by step, and participants could try it out themselves. During the session, the evaluation team took notes on observations of the problems encountered and usability problems experienced by the participants and documented their suggestions for improvements. After the workshop, the participants answered a structured questionnaire.

We asked the workshop participants to provide an overall estimation of the different InTaVia components on a scale of

[32] Accessible at https://intavia.acdh-dev.oeaw.ac.at
[33] https://tinyurl.com/InTaViaExample1
[34] https://tinyurl.com/InTaViaExample2
[35] The API is documented at https://intavia-backend.acdh-dev.oeaw.ac.at/v2/docs using OpenAPI 3 specification of parameters. The source code is available at https://github.com/InTaVia/InTaVia-Backend.
[36] see https://github.com/InTaVia/InTaVia-Backend for implementation details.
[37] The API is available at https://intavia-backend.acdh-dev.oeaw.ac.at/v1/recon/reconcile/.
[38] https://github.com/InTaVia/InTaVia-Backend



(a) search and curation result screen **(b)** storytelling suite example

Figure 2. Two example screenshots of the InTaVia platform

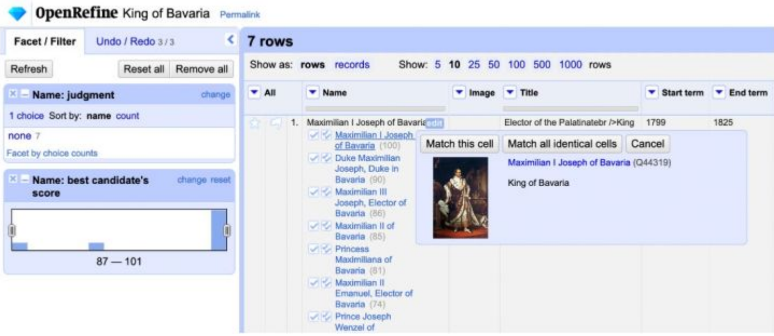


Figure 3. Example use of the IKG data in OpenRefine through the reconciliation API

1 (positive) to 7 (negative). All components were evaluated above average, with the IKG itself receiving the second-highest score of 1.8 after the visual analysis components. Regarding the IKG, the workshop participants found it sound (3 respondents) and liked the “prosopographical aspect”, the “ability to compare and join data from different persons, their relations” as well as the ”internal references possibilities” and that ”it is connecting data from different sources”. They also highlighted that it is “open source” and their appreciation of ”front-end simplicity”.

The problems encountered during the workshops were mainly related to usability for participants with lower levels of prior knowledge, who found that “it takes some time to learn how to use it”, “it is a little bit complicated”, and that they made “mistakes :)”. With respect to the IKG they criticized a “lack of connectivity of Slovenian archives” and “closed metadata sets (meaning low visibility) and option to use all of the above”. Finally, they missed “support for certain web browsers”. Suggestions for improvements included “more introductory documentation to walk me through the tools/ functionalities “, “support for multiple browsers”, “connecting to Eastern Europe archives, libraries”.

The workshop participants offered nuanced and largely positive feedback on the IKG, the InTaVia platform and its modules. They clearly identified ways it could be integrated into their existing user practices, as well as new and relevant practices emerging from this innovative approach to linking, visualizing, and communicating cultural heritage data.

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.

Note that in our approach, entities are not merged together, but *linked* via a provided entity in a dedicated graph. Using that, the API and Frontend UI are able to provide a holistic view. E.g. for ”Giuseppe Acerbi” the UI shows a stream of events that come from the different datasets[39]. There are e.g. two birth events (from Finnish and Austrian data) and CH data from Wikidata. The same systematic can be used in SPARQL[40].

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

[39]See `https://tinyurl.com/InTaViaExample3`
[40]See `https://qlever-ui.acdh-dev.oeaw.ac.at/intavia/W8TINE`

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] Dijkshoorn C, Aroyo L, van Ossenbruggen J et al. Modeling cultural heritage data for online publication. *Applied Ontology* 2018; 13(4): 255–271.

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

[3] Szekely P, Knoblock CA, Yang F et al. Connecting the Smithsonian American Art Museum to the linked data cloud. In *The Semantic Web: Semantics and Big Data*. Springer Berlin Heidelberg, pp. 593–607. DOI: 10.1007/978-3-642-38288-8_40.

[4] de Boer V, Wielemaker J, van Gent J et al. Amsterdam Museum linked open data. *Semantic Web* 2013; 4(3): 237–243.

[5] Hyvönen E, Mäkelä E, Kauppinen T 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.

[6] Mäkelä E, Hypén K and Hyvönen E. BookSampo—lessons learned in creating a semantic portal for fiction literature. In *International Semantic Web Conference*. Springer, pp. 173–188.

[7] Koho M, Ikkala E, Leskinen P et al. WarSampo knowledge graph: Finland in the second world war as linked open data. *Semantic Web* 2021; 12(2): 265–278. DOI:10.3233/SW-200392.

[8] Leskinen P and Hyvönen E. Reconciling and using historical person registers as linked open data in the AcademySampo portal and data service. In *International Semantic Web Conference*. Springer, pp. 714–730.

[9] de Boer V, van Rossum M, Leinenga J et al. 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*. Springer, pp. 229–244.

[10] Haslhofer B and Isaac A. data.europeana.eu: The Europeana linked open data pilot. In *International conference on Dublin Core and metadata applications*. pp. 94–104.

[11] Schlögl M and 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*. NAP, New Academic Press, 2020. pp. 31–48. URL https://www.oeaw.ac.at/fileadmin/Institute/ACDH/OEBL/pdf/_apis_Buch_WEB.pdf.

[12] TEI Consortium (ed.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium, 2025. URL http://www.tei-c.org/Guidelines/P5/. Accessed: 12 December 2025.

[13] Erjavec T, Dokler J and Ogrin PV. Slovenian biography. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*. CEUR Workshop Proceedings, pp. 16–21. URL https://ceur-ws.org/Vol-2119/paper3.pdf.

[14] Fokkens A, Ter Braake S, Ockeloen N et al. BiographyNet: Extracting relations between people and events. *arXiv preprint arXiv:180107073* 2018; .

[15] Hyvönen E, Leskinen P, Tamper M et al. BiographySampo - publishing and enriching biographies on the semantic web for digital humanities research. In *The Semantic Web. ESWC 2019*. Springer-Verlag, pp. 574–589. DOI:10.1007/978-3-030-21348-0_37.

[16] Schlögl M, Windhager F, Mayr E et al. Biographische Informationssysteme (DPBs, Digital Knowledge Databases, Virtual Research Environments), 2019. DOI:10.5281/zenodo.2593761.

[17] Ockeloen N, Fokkens A, Ter Braake S et al. BiographyNet: Managing provenance at multiple levels and from different perspectives. In *LISC@ ISWC*. pp. 59–71.

[18] Lagoze C, Van de Sompel H, Nelson ML et al. Object re-use & exchange: A resource-centric approach. *arXiv preprint arXiv:08042273* 2008; .

[19] Doerr M, Gradmann S, Hennicke S et al. The Europeana data model (EDM). In *World Library and Information Congress: 76th IFLA general conference and assembly*, volume 10. p. 15.

[20] Doerr M. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine* 2003; 24(3): 75–92.

[21] Tuominen J, Hyvönen E and 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)*. CEUR Workshop Proceedings, pp. 59–66. URL https://ceur-ws.org/Vol-2119/paper10.pdf.

[22] Lebo T, Sahoo S and McGuinness D. PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013, 2013. URL https://www.w3.org/TR/2013/REC-prov-o-20130430/.

[23] Knublauch H and Kontokostas D. Shapes constraint language (SHACL). Technical report, W3C, 2017. URL https://www.w3.org/TR/shacl/.

[24] de Boer V, Wielemaker J, van Gent J et al. Supporting linked data production for cultural heritage institutes: The Amsterdam Museum case study. In Simperl E, Cimiano P, Polleres A et al. (eds.) *The Semantic Web: Research and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-30284-8, pp. 733–747.

[25] Kusnick J, Mayr E, Seirafi K et al. Every thing can be a hero! narrative visualization of person, object, and other biographies. *Informatics* 2024; 11(2). DOI: 10.3390/informatics11020026. URL https://www.mdpi.com/2227-9709/11/2/26.

[26] Delpeuch A, Pohl A, Steeg F et al. Reconciliation Service API v0.2: A protocol for data matching on the web. Final Community Group Report 10 April 2023, W3C, 2023. URL https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/.

[27] project I. Final user evaluation report deliverable 7.3, 2023.