9th International Conference on Prestigious Applications of Intelligent Systems – PAIS@ECAI2020
Santiago de Compostela, Spain

Cross-Border Medical Research using Multi-Layered and Distributed Knowledge

Gábor Bella[1], Liz Elliot[2], Subhashis Das[3], Stephen Pavis[4], Ettore Turra[5], David Robertson[6], and Fausto Giunchiglia[7]

Abstract. As medical research becomes ever finer-grained, experiments require healthcare data in quantities that single countries cannot provide. Cross-jurisdictional data collection remains, however, extremely challenging due to the diverging legal, professional, linguistic, normative, and technological contexts of the participating countries. Medical data heterogeneity, in particular, is still a largely unsolved problem on the international level, due to the complexity of data combined with strict precision and data protection constraints. We propose a scalable solution based on a novel knowledge architecture and the corresponding knowledge graph integration methodology. Medical knowledge that drives the scalable integration process is divided into multiple functional layers and is maintained in a distributed manner across participating countries. We successfully applied the approach in the context of a research experiment across Scotland and Italy, and are currently adapting it within other initiatives of Europe-wide health data interoperability.

1 INTRODUCTION

Advances in medical research have led to data-intensive methods such as *stratified* and *precision medicine* that determine fine-grained diagnoses and therapies for specific types of patients. Such methods, however, require patient data to be collected from ever larger pools of population in order to reach statistical significance. Despite high public and private demand for large amounts of research data, the overall difficulty and cost of reaching beyond populations of single jurisdictions remain prohibitive for technological, legal, and economic reasons.

Our paper specifically addresses *data heterogeneity*, a particularly challenging problem in the context of international medical research. Beyond the obvious (nonetheless hard) problem of working across languages, data interoperability on an international level also needs to align local standards and practices that may currently be codified at varying degrees of formality. The solution must also scale with respect both to the amount of data and the number of participating countries. Finally, the solution must respect severe constraints on high data accuracy and local privacy rules with respect to sensitive medical data.

In the context of a research experiment initiated by the *National Health Services of Scotland* and conducted between Scotland and the Italian province of *Trento*, we tackle cross-border data integration through a solution based on the *integration of local and international medical knowledge*. Following a divide-and-conquer approach, integrated knowledge is *multi-layered*, with each layer addressing heterogeneity on a different representational level: *natural language*, *terminology*, and *schemas*. Furthermore, it is *distributed* into *local* and *international* knowledge bases where each such knowledge instance addresses a specific subproblem of data integration, namely the formalisation of local data as a knowledge graph and the subsequent mapping of local graphs across borders. The knowledge-driven solution we propose is privacy-aware by design as, contrary to data-driven AI approaches, it does not transfer any data across jurisdictions for the purposes of integration, in respect of local laws and regulations [17].

[1] DISI, University of Trento, Italy, gabor.bella@unitn.it
[2] Usher Institute, University of Edinburgh, UK, elizabeth.elliot@ed.ac.uk
[3] DISI, University of Trento, Italy, subhashis.das@unitn.it
[4] NHS Scotland, UK, stephen.pavis@nes.scot.nhs.uk
[5] APSS di Trento, Italy, ettore.turra@apss.tn.it
[6] School of Informatics, University of Edinburgh, UK, d.robertson@ed.ac.uk
[7] DISI, University of Trento, Italy, fausto.giunchiglia@unitn.it

2 STATE OF THE ART

Among AI-based approaches to combining health data for research, recent data-driven solutions rely on machine-learning-based analytics, such as *Watson Health* [10]. While efficient where the data does not present deep problems of semantic heterogeneity—as in medical image analysis [12], or when the system is trained and used within the same context, such as the same hospital or country—they have been found much less robust when transferred from one country to another, for lack of addressing semantic heterogeneity in an explicit and pervasive manner [19]. The lack of explainability of decisions taken by learning algorithms is also a dissuading factor for the medical community.

The standard approach to solving healthcare data heterogeneity problems expects local data controllers to map their data to a *common data model* (or 'shared ontology' in Semantic Web terms), such as HL7 or, more recently, OMOP and FHIR [18] on the schema level and to standard terminologies such as SNOMED or ICD on the data value level. While standard data models and terminologies do address heterogeneity and are a necessary element of interoperable solutions, in themselves they are not sufficient: the methodology through which they are applied determines in a large part whether the overall solution can scale with new countries, data providers, and data records. Thus, solutions that leave the burden of translation and alignment to international standards on local data controllers (hospitals, labs, etc.), such as in [16], cannot scale, due to the necessity of (constantly evolving) linguistic, standards-related, medical, and technological expertise. Such an effort is outside the competences and means of local data controllers, which explains the lack of large-scale cross-border research.

The conventional solution for conversion to common data models

is to implement ad-hoc procedural logic, as in [20] or [15]. Procedural approaches, however, are onerous and inflexible on the long term as the evolution of experiment requirements, data sources, or the target representation inevitably leads to new software development cycles. In order to reduce such costs, data cleaning, mapping, and ETL tools using formal knowledge (e.g. OWL ontologies) have been proposed, such as the domain-agnostic *Karma* [11]. These tools successfully tackle schema-level heterogeneity—and we reuse them for this purpose in our work—but have no support for multilingual data or the mapping of large domain terminologies.

Ontology-driven approaches to biomedical data integration have been proposed as a solution that reduces the cost of interoperability, as shown in a recent survey [5]. *UMLS* has been the most complete and well-known aggregator of multilingual medical standards [4]. While it represents medical knowledge on multiple levels (lexical, terminological, ontological), its largest component is the *Metathesaurus* that integrates biomedical terminologies. While UMLS and similar aggregators are used in research projects as sources of aligned terminological knowledge [1, 5, 14], they cannot cover *local knowledge* in sufficient detail to support a deep and comprehensive interpretation of local datasets. Also, they are not under the control of experimenters or data controllers, and are inflexible and slow to adapt to new research problems, participating countries, standards, or to constantly changing data representations.

3 CHALLENGES OF CROSS-BORDER RESEARCH: THE *DOAC* USE CASE

The research experiment introduced below, initiated and overseen by Scottish medical researchers, highlights the major technical difficulties in cross-border data sharing. The goal of the research was to examine, for patients having had an *intracranial hæmorrhage*, the safety of taking a specific category of drugs called *direct oral anticoagulants* (DOAC) with respect to more traditional treatment (i.e. *Warfarin*). As the experiment could not provide statistically significant results on data from the *National Health Services of Scotland (NSS)* alone, due to the relative novelty of DOACs and the specificity of medical conditions examined, the study also integrated data from Italy, specifically from the *Azienda Provinciale per i Servizi Sanitari (APSS)* of the province of Trento.

Cross-border experimentation, however it may be realised in practice, needs to traverse the following macro-steps: (1) *experiment definition* that reaches a common prior agreement among data controllers in each country with respect to the experiment requirements, down to the dataset and attribute level; (2) *data extraction:* each data controller extracts from local databases all data necessary for the experiment; and (3) *data conversion:* the conversion of extracted data to the representation required for the experimentation.

While these steps are not specific to cross-border settings, their complexity is greatly increased in such scenarios. Our survey, involving NHS Scotland data analysts, showed that even in a single-country scenario, for each individual experiment, the entire pre-analytics data provision process typically lasts between 6–24 months (!), with actual data preparation taking up to one month of actual work and up to 10 months of elapsed time. As we show below, the considerably higher complexity of cross-border data heterogeneity is one of the main reasons for this situation.

Language-level heterogeneity. The use of natural language is pervasive through each of the macro-steps above. In step 1, the experiment description needs to be understood and evaluated by data controllers in each participating country. In step 2, datasets, data attributes, and data values need to be queried and filtered in the local language. For example, in the Italian drug product database, the value *'10 compresse rivestite'* (meaning *'10 coated tablets'*) contains both the quantity of items contained in a pack and the nature of the item, either of which may be needed to be extracted for the experiment. The same description appears also as *'compresse riv.'* and as *'compresse rivestite con film'*, which need to be understood as equivalent. Finally, in step 3, data expressed in the local languages need to be combined and then translated to the language in which the experiment will be carried out, which may or may not be among the languages of the originating countries.

Terminology-level heterogeneity. Beyond natural language, a major difficulty resides in understanding the precise meaning of medical terms or codes used within experiment descriptions, schemas, and data values. For example, the Scottish prescription attribute *'Quantity prescribed'* refers to the number of tablets while its Italian counterpart *'Quantità'* to the number of packs. Within data values, local, national, and international codes are frequently used within medical data and are essential for automating large-scale research. Our experiment involved disease codes, medical procedure codes, as well as local codes for *admission type*, *admission reason*, and socio-economic *status of deprivation*. To encode diseases, Scotland uses the *ICD-10* international standard[8] while Italy its previous version *ICD-9* including, however, *Clinical Modifications* originating from the United States. Procedures are encoded in Scotland using the national *OPCS-4* standard[9] while Italy uses *ICD-9-PCS*.

Schema-level heterogeneity. Each local data controller—hospital, lab, or regional/national controller—uses its own data structures for representing patients, prescriptions, visits, etc. The state-of-the-art approach is (1) to understand in advance which local datasets will need to be used; (2) to map these local schemas to a standard pivot schema, such as OMOP or FHIR, using ad-hoc data conversion logic at design time; and (3) to run data extraction and conversion automatically during experiment preparation. While this approach is straightforward on paper, in practice it often proves to be costly to maintain: firstly, the local and the pivot schemas evolve over time, which the conversion implementation needs to follow. Secondly, new research experiments may require new and unforeseen datasets, which again means extending the conversion logic. Thirdly, data heterogeneity is often a problem even internally to jurisdictions (e.g. Italian regions have a high degree of freedom in managing their data), which makes the implementation of cross-border mapping subject to solving internal data integration problems first.

Constraint of distributedness. Several factors make fully centralised data integration approaches unsuitable. For data protection reasons, many jurisdictions—such as the UK in our experiment—do not allow sensitive medical data to leave the country of origin. This constraint imposes a system architecture that performs integration in a distributed manner. Competences are also distributed: the extraction of data from legacy DBs and the deep understanding of such data can only be done by local data controllers that are familiar with the local language, standards, and practices. Data conversion, on the other hand, requires an understanding of the shared data formats, language, terminologies, and schemas. Local and shared data representations evolve independently, the former due to changing operational healthcare needs while the latter driven by medical research.

[8] The International Classification of Diseases, 10th rev.
[9] OPCS Classification of Interventions and Procedures version 4: https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/10

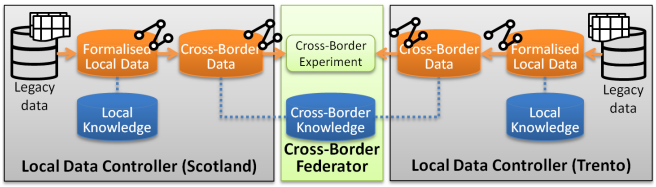

Figure 1. The distributed knowledge and data integration architecture.

4 MULTI-LAYERED AND DISTRIBUTED KNOWLEDGE

We tackle the challenges above through *multilingual, knowledge-driven, and distributed data integration*. The distributed architecture consists of *local and cross-border knowledge*, as shown in Figure 1. *Local knowledge bases* are under the full control of the local data controllers participating to the cross-border experiment, such as a hospital or a higher-level—regional or national—government institution. Their role is to *formalise* local data, i.e. describe them in a fine-grained and unambiguous way. In turn, the *cross-border knowledge base*, under the responsibility of a *Cross-Border Federator* entity, has the role of maintaining mutually agreed *shared* medical knowledge and help in performing mappings from local to shared representations. Decoupling the *formalisation* of local data from its *mapping* results in: (1) better maintainability as local and shared representations may evolve independently; (2) a simplification of the mapping logic as it operates over a non-ambiguous formal representation instead of raw data, leading to easier automation and thus better scalability; and (3) a natural separation of tasks along competences: formalisation is in the hands of local data experts while mapping is performed by interoperability experts.

Each knowledge base is organised into three interconnected layers defining linguistic, terminological, and ontological medical knowledge, each layer tackling a specific form of data heterogeneity. This layered structure provides compositionality to knowledge bases, making them easier to adapt to local needs.

The resulting knowledge architecture is thus divided into three horizontal layers and is distributed among vertical knowledge instances: in our experiment, two local ones and a cross-border one. The three layers are built upon each other and, likewise, the knowledge instances are interconnected in a hierarchical manner, realising an integrated knowledge architecture as depicted in Figure 2.

4.1 The language layer

The language layer describes the words and expressions used within medical data: general and domain terms, data attribute names, as well as medical codes that we consider as part of medical language. It is built from three principal sources: (1) the general lexicon, as it is frequently used even in domain applications: we use *wordnets* of the given languages[10] [13]; (2) medical domain terminologies such as SNOMED CT, ICD, or LOINC; and (3) other labels and terms relevant with respect to the research experiment, encoded manually by the maintainers of the knowledge instance.

Within local knowledge, the language layer encodes words and terms in the local language and with local relevance. For example, in Figure 2, the attribute name *'mpridia'* and the term *'diagnosi primaria'*, both standing for *primary diagnosis*, are included in the language layer of the Trento local knowledge. Likewise, the UK-specific OPCS-4 procedure code *'X39.1'* meaning *'oral administration of therapeutic substance'* is encoded as part of the local language.

Within cross-border knowledge, the role of the language layer is to describe experiment data in human-readable form, in the language of the experimenter. While in our case this language was English, simultaneous support of multiple languages is possible by plugging in multiple language layers.

4.2 The terminology layer

The terminology layer is composed of *language-independent concepts* that represent the meanings of natural language words, domain terms, or schema labels from the language layer, and *concept relations* that organise concepts into a graph, such as *is-a* and *part-of*. For example, in Figure 2, *'hæmorrhage'*, *'hemorrhage'*, and *'R58.X'* are linked to the same concept and therefore are considered as synonyms. The concept of *'intracranial hæmorrhage'*, in turn, is related to the concept of *'hæmorrhage'* through subsumption.

Local terminology contains concepts of local relevance, representing locally used terms and schemas. Cross-border concepts, in turn, are generated from international standards (e.g. ICD or SNOMED CT) as a way to guarantee the highest possible level of interoperability.

The terminology layer is a key device for tackling the heterogeneity of languages and medical standards alike. On the local level, through a process of *formalisation*, it maps informal language-based data representations (text within data, coded values) into a formal graph-based representation where both schema and data elements are represented as language-independent concepts. Local knowledge can also be used for solving local heterogeneity issues. On the cross-border level, the terminology layer acts as a bridge both across languages and standards. Equivalence mapping relations between local and cross-border concepts provide 'cross-walks' both across languages and across local and international standards. For example, in Figure 2, such relations assert the equivalence of the Italian concept labelled as *'diagnosi primaria'* with the Scottish concept of *'main condition'* through mapping both to the pivot SNOMED concept of *'main diagnosis'*. Likewise, the concept labelled as *'4590'* from the ICD-9 standard currently in use in Italy is mapped as equivalent to the ICD-10 concept of *'R58.X'*, ICD-10 being the pivot standard in the cross-border instance.

We build concepts and concept relations from wordnet *synsets* and *synset relations* for general language, and from *terminological units* and *term relations* for domain terminologies. While domain terminologies may overlap with the wordnet and also among themselves (e.g. both ICD and SNOMED define *'cerebral hæmorrhage'*), this rarely poses a problem in practice as medical schemas usually provide context on which terminology is being used within a given data attribute (e.g. the *main condition* attribute always uses ICD). In order to favour the correctness and maintainability of knowledge, we adopt a lightweight approach to integrating wordnets and terminologies: a domain expert manually combines them simply by attaching the root terms of specialised terminologies to subsumer terms within wordnets or higher-level reference terminologies, via *is-a* relations. For example, the root concept of the ICD disease hierarchy is subsumed by the wordnet-sourced concept of *'pathological state'*.

Cross-lingual and cross-standard mappings are provided by high-quality expert-curated resources: for general language concepts, we

[10] Wordnets are obtainable for a large number of languages from http://globalwordnet.org/wordnets-in-the-world/.

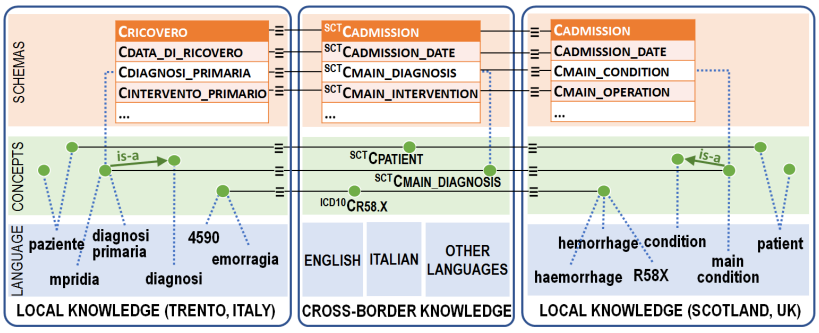

Figure 2. The three-layered knowledge structure, divided across two local and one cross-border instance.

reuse existing mappings towards the pivot Princeton WordNet English synsets provided by most wordnets, as described in [7]. For medical concepts, cross-lingual mappings are provided by multilingual domain terminologies (such as SNOMED CT, ICD, or LOINC). Cross-standard equivalence mappings between local and pivot concepts (such as from ICD-9 used locally in Italy to ICD-10 used internationally) originate either from crosswalks hand-curated by the international medical community, or are added by local data controllers. The more extensively a local controller adopts well-known standards for local use, the smaller the manual effort needed to create mappings. In any case, curating the mappings used by the system is entirely under the control of experts.

4.3 The schema layer

The schema layer, described formally in our earlier work [6, 8], models data structures (data schemas, ontologies) in an entity-centric manner. Aggregating attributes around well-known domain entities (patient, admission, drug, prescription, etc.) as opposed to, e.g. more normalised or triple-based schemas increases the cross-border understandability of data representation and, hence, interoperability. This is why international schema-level health interoperability standards, such as FHIR or OMOP, also follow an entity-centric approach.

Schemas, attributes, and attribute datatypes are defined via concepts from the terminology layer. In the example of Figure 2, the concept of the schema is $C_{\text{Admission}}$, the concept of its attribute is $C_{\text{main\_diagnosis}}$, and the concept of the attribute datatype is $C_{\text{pathological\_state}}$, meaning that its values are concepts subsumed by it. The fact that schema elements are defined through unambiguous formal concepts as opposed to natural-language labels facilitates the definition of cross-border schema mappings and also allows schemas to be displayed in any language that experimenters wish to use and that are supported by the language layer of cross-border knowledge.

While local schemas are defined via local concepts, schemas defined within cross-border knowledge use standard schemas for interoperability: in our setup the schemas were defined on the basis of OMOP, designed specifically for medical research.

5 KNOWLEDGE AND DATA INTEGRATION METHODOLOGY

This section presents the methodology on the use of multi-layered and distributed knowledge for enabling cross-border integration of data. We use the DOAC experiment as an example case study to validate the approach.

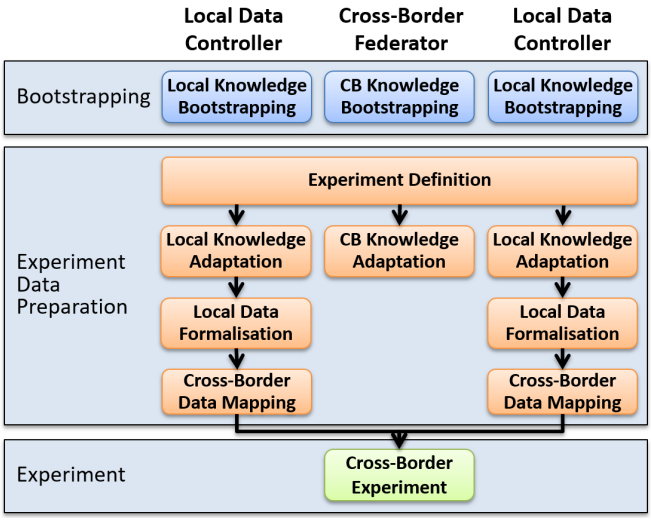

Figure 3. High-level steps of the experiment data preparation process. While *cross-border data mapping* may take place within local jurisdictions for privacy reasons, it remains governed by cross-border knowledge.

The high-level steps of the experiment data preparation process, as depicted in Figure 3, are articulated as follows:

1. cross-border experiment definition;
2. local and cross-border knowledge adaptation;
3. knowledge-based data integration, consisting of local data formalisation followed by cross-border data mapping.

Figure 3 also shows that the experiment-specific process above is preceded by a one-time bootstrapping effort that sets up the cross-border knowledge base as well as local knowledge bases in each participating jurisdiction. Both on the local and the cross-border levels, the data preparation process is overseen by *data scientists* who have a deep understanding of the contents of the medical datasets used, as well as of domain knowledge and the best practices of knowledge representation. In the experiment definition step, however, the active participation of medical researchers is also necessary.

Below we provide details on how the DOAC experiment preparation steps were implemented using our knowledge architecture. For the storage, creation, maintenance, and querying of multi-layered knowledge instances we used a knowledge base technology developed at the University of Trento, described in [9]. While a full account of the automation of data integration is beyond the scope of this paper, in sections 5.4 and 5.5 we provide a summary on the tools and approaches used.

5.1 One-time knowledge bootstrapping

Knowledge bootstrapping is a one-time initial effort to create the knowledge bases that will drive data integration: one for each local data controller and one for the cross-border integrator. At this stage the most frequently used knowledge is set up, such as the general vocabulary, disease and procedure codes, and the core data schemas. It is not needed to achieve an *a priori* exhaustive coverage of medical data: knowledge bootstrapping is only a first step of a long-term iterative medical knowledge management process. In subsequent experiments this bootstrapping step is omitted.

In the DOAC experiment three knowledge bases were used: two local ones for Scotland and Trento, as well as a cross-border knowledge instance installed in Scotland. The knowledge resources used for bootstrapping, including their sizes, are shown in Table 1. Local knowledge was filled with locally used encoding schemes and data labels. As shown in Table 1, English and Italian wordnets were preloaded to cover the general lexicon, and for disease codes we loaded ICD-10 for Scotland and ICD-9-CM for Italy. Cross-border knowledge was populated with international standards: SNOMED CT as a backbone reference ontology of concepts and terms, as well as the ICD-10-CM hierarchy. These domain terminologies were attached by their root to concepts of the general lexicon by a data scientist. Pre-loaded knowledge also included expert-curated mappings across standards provided by the *US National Library of Medicine*, NSS, and APSS.

5.2 Experiment definition

As described in section 3, experiment definition is one of the most time-consuming phases of cross-border experimentation, as a common understanding and agreement needs to be reached among participating humans in a context of multilingual and heterogeneous data. Integrated formal knowledge accelerates this process by allowing immediate visibility to a formal and uniform description of datasets across jurisdictions to both experimenters and data controllers.

As our pilot DOAC experiment was conducted in parallel with the bootstrapping of knowledge, we did not benefit of this advantage and had to define the experiment based on human expert knowledge alone. The initial Scottish experiment proposal was first reviewed by the APSS data controllers and further precisions were requested with respect to the meaning of UK-specific variables and the level of formality in describing the cohort. Due to the physical distance and language issues, communication was mediated by the knowledge and data integration team. Once mutual understanding was reached about the experiment goals and criteria, APSS reviewed each data attribute requested by Scotland and evaluated equivalences, or lack thereof, with Italian attributes. A number of attributes were signalled either as having no equivalent in Italy or not being retrievable: this was the case of *admission reason*, *care home residency flag*, *cause of death*, and *measure of deprivation*. On the basis of this input, in the final experiment definition, these variables were either dropped or, in the case of *cause of death*, were approximated by hospital re-admission disease codes that were deemed by experts to be strongly linked to the cause of death.

5.3 Knowledge adaptation

Based on the experiment definition, the local and cross-border knowledge instances were adapted to the experiment by a data scientist using interactive knowledge management tools for the creation

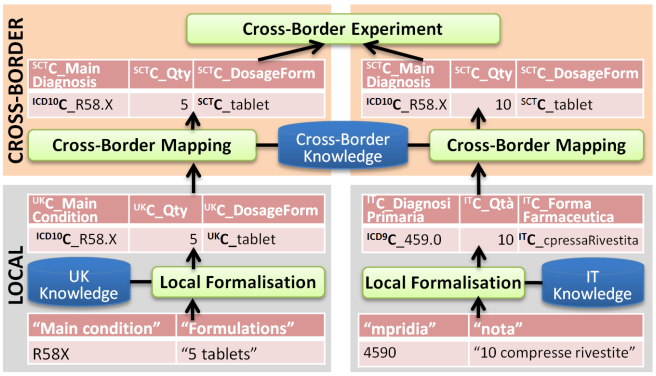

Figure 4. Illustration of the two steps of data integration: local formalisation (bottom layer) followed by cross-border mapping (top layer). The two columns represent a Scottish and an Italian jurisdiction.

of new lexical items, concepts, schemas, and attributes. As the data schemas used (see Table 1) depend on the experiment, they were created in the adaptation phase rather than during bootstrapping. For the design of cross-border schemas and attributes, the OMOP research data exchange standard was used as a general reference, but the actual schemas were limited to the needs of the experiment, following our general philosophy of extending cross-border knowledge on an on-demand basis. The attributes of cross-border schemas were defined as SNOMED CT concepts whenever available, while local ones were represented by *ad hoc* concepts.

Still as part of knowledge adaptation, relevant DOAC drugs (*Rivaroxaban, Dabigatran, Apixaban,* as well as *Warfarin* for control) and the corresponding drug products were imported were defined in all three scopes, as well as the creation of pivot OMOP-based entity-centric schemas.

5.4 Local data formalisation

Local data formalisation (bottom half of Figure 4) converts data that is informally or semi-formally expressed, typically as relational tables containing textual and numerical values, into a formal *knowledge graph*:

- legacy (typically relational) data schemas are formalised by mapping their attributes to the formal, entity-centric, language-independent schemas defined in the local knowledge;
- data records become *entities*, instances of the schemas above;
- medical codes and relevant natural-language terms within attribute values are formalised as concepts: either single concepts, e.g. *'tablet'* $\rightarrow \mathrm{C}_{\mathrm{Tablet}}$) or *'4590'* $\rightarrow \mathrm{C}_{\mathrm{ICD9:459.0/Hæmorrhage}}$, or short phrases containing multiple concepts, e.g. *'5mg coated tablets'* $\rightarrow <5, \mathrm{C}_{\mathrm{mg}}, \mathrm{C}_{\mathrm{CoatedTablet}}>$;
- foreign keys, numerical IDs, named entities, and other such identifiers within data are formalised as entity links;
- other (e.g. numerical) values are formalised as simple datatypes.

The resulting knowledge graph is thus constituted of *entities*, *concepts*, and *datatypes* as nodes, and *entity-to-entity*, *entity-to-concept*, and *entity-to-datatype* links as edges.

For both Scottish and Italian data, we carried out the mapping of schemas and the underlying data records semi-automatically using *StarLinker*, a semantic data transformation tool that extends *Karma* [11]. The main difference of StarLinker with respect to Karma and

Table 1. Summary of the contents of the three knowledge bases used in the case study, including knowledge sizes in terms of number of labels (for terminologies), attributes (for schemas), and relations (for mappings).

Resource	Scotland	Size	Italy	Size	Cross-border	Size
general lexicon	Princeton WordNet	120k labels	Italian MultiWordNet	35k labels	Princeton WordNet	120k labels
ref. terminology	n/a		n/a		SNOMED CT Intl	1.2M labels
diseases	ICD-10	40k labels	ICD-9-CM	140k labels	ICD-10-CM	70k labels
drugs	BNF (part)	6 labels	AIC (part)	6 labels	SNOMED CT	12 labels
experiment data schemas	SMR01/admissions, PIS/prescriptions, NRS death records	77 attrs 50 attrs 52 attrs	Ricoveri, Prescrizioni	21 attrs 10 attrs	OMOP:Visit_occurrence, OMOP:Dose_era, OMOP:Death	13 attrs 9 attrs 8 attrs
terminology mappings	BNF ↔ SNOMED CT	6 rels	MWN ↔ PrincetonWN, ICD9CM ↔ ICD10CM, AIC ↔ SNOMED CT	31k rels 23k rels 5 rels	SNOMED ↔ ICD10CM	21k rels
schema mappings	SMR01 → Visit_occ., PIS → Dose_era, NRS DR → Death	9 rels 12 rels 7 rels	Ricoveri → Visit_occ., Prescrizioni → Dose_era, Ricoveri → Death	5 rels 17 rels 2 rels	n/a	

other ETL tools that carry out schema mappings and data transformations is the ability to disambiguate natural language text and medical codes contained in data into unambiguous concepts. StarLinker relies on the *SCROLL* multilingual NLP tool [3] and the terminology layer of local knowledge to perform language-independent word sense disambiguation on labels in structured data (the method is described in [2, 3]), and can reach a high precision due to the constrained nature of the disambiguation tasks (e.g. in prescription data, it only needed to extract concepts such as *mg* or *capsule* where disambiguation is constrained by the subsumer SNOMED concepts of *unit of measure* and *dosage form*, respectively).

In order to ensure the precision of data formalisation, which is of particular importance for medical research, the initial definition of all formalisation steps through StarLinker is overseen and validated by a local data scientist. StarLinker then allows the fully automated application of the same processing steps over subsequent datasets, which is a key feature for scalability over large amounts of data (we provide a scalability experiment in the next section, Table 4).

On the Scottish side, three datasets were formalised: *Inpatient admissions (SMR01)*, *Prescriptions (PIS)*, and *NRS Death Records*. Within admission data, diagnosis codes were identified using the ICD-10 standard represented in the Scottish knowledge base. In prescriptions, local *BNF (British National Formulary)* item codes were referring to official textual drug descriptions (e.g. *'5mg tablets; 10 tablets'*) that we had to parse to extract drug names and prescribed doses. Dates and causes of death, the latter as ICD-10 codes, were extracted from the *NRS Death Records* dataset.

On the Italian side, in the *Ricoveri* (inpatient admissions) dataset, diagnoses expressed as ICD-9-CM codes were recognised and labelled with the corresponding concepts using the Trento knowledge base. For causes and dates of deaths, as it was not possible to obtain Italian death records for the time period defined in the experiment, hospital readmission data was used, including diagnosis (ICD) codes. Italian prescriptions contained coded drug identifiers in the local AIC format, that we used to retrieve dosages and prescription models (e.g. *'orale 0.3 grammi'*, *'5 mg 60 compresse rigide'*). From this dataset, drug names, quantities, and units of measure were extracted, normalised, and labelled with meaning.

In total, formalisation involved the definition of 37 unique data transformation rules (e.g. for the canonical formatting of dates and medical codes, or the splitting of columns containing complex values) and 10 unique concept extraction rules (disease codes, sex, units of measure, dosage forms, drugs, routes of administration, etc.).

5.5 Cross-border data mapping

Cross-border data mapping (top half of Figure 4) takes a formal yet still locally specific knowledge graph as input and maps it to the cross-border representation. Due to data protection constraints, data mapping typically happens independently at each data controller, using shared cross-border knowledge.

We mapped local Scottish and Italian schemas to the pivot OMOP-based schemas semi-automatically, using the StarLinker tool. The mapping of concepts across terminologies (see Table 1 for details) was fully automatic whenever expert-curated one-to-one equivalence mapping rules were available, otherwise it required the manual extension of knowledge. When mapping Italian ICD-9 codes to cross-border ICD-10, 16 codes out of the total 23 appearing in the Italian cohort could not be automatically mapped due to the lack of equivalence mappings across the two standards. These mappings needed to be added manually by a data scientist based on input from a researcher. In general, for strongly context-specific or fine-grained information, no mapping may be possible without some level of information loss; however, such losses are detectable and manageable prior to running the research experiment, thanks to the formal representation. For example, in Figure 4 the Italian concept of *'compressa rivestita'*, meaning *'coated tablet'*, is mapped to the broader concept of *tablet* in case the pivot format does not represent the former: while some detail is lost, the information remains formally correct.

In total, the two final mapped knowledge graphs consisted of 87,165 triples computed from 7,717 initial data records, not counting the triples describing the knowledge underlying the process.

6 LESSONS LEARNT

Precision and completeness. A specificity of data integration in the medical domain is the absolute need for precision, i.e. no information lost or corrupted during processing. The possibility of human supervision and the ability to explain automatically obtained results to human experts are strong requirements of the medical community. The knowledge-based and semi-automated method, overseen by a data scientist, offers a workable solution: knowledge mappings, while never complete, are transparent, immediately visible, and extensible. The fact that the Karma-based StarLinker data processing tool is able to work in both semi-automated and fully automated modes provides the possibility to the medical community to balance the need for supervision with scalability at their will.

We conducted an experiment on the precision and coverage of the semi-automated components of the data formalisation and cross-border mapping process. Three such components were evaluated: (1) procedural data transformations to reduce heterogeneity in dates or natural language labels; (2) concept extraction (that converts natural language labels into concepts); and (3) knowledge-based concept mapping. The latter two were evaluated both after their initial fully-automatic execution and after expert curation of the automatic results. Curation included both reviewing the concept selected by the system from multiple alternatives, and extending the knowledge in case of missing concepts. The results in Table 2 show that precision was always very high, which is due to the relative simplicity of labels as well as word sense disambiguation having been carefully constrained by attribute-specific concepts. Initial concept and especially mapping coverage were low, which was a knowledge incompleteness problem. Once the completeness issue addressed by knowledge adaptation, precision and coverage could be increased to 100%. Finally, data transformations provided perfect results due to the original data being regular with little noise.

Table 2. Size, precision, and coverage statistics of the semi-automated data mapping process using StarLinker.

Task	Nb	Prec.	Coverage
unique concepts extracted (auto)	63	93.7%	68.2%
unique concepts extracted (curated)	63	96.8%	100%
unique concept mappings (auto)	23	100%	26.1%
unique concept mappings (curated)	23	100%	100%
unique data transformation rules	37	100%	100%

Scalability. In order to understand the cost of the proposed method in terms of time, in Table 3 we provide information of the effort spent in the various steps of data preparation. 55% of the total time went into knowledge bootstrapping, 27% into experiment definition, only 9% into actual experiment-specific knowledge and data preparation, and another 9% into the manual validation of transformed data. Apart from bootstrapping, not to be redone for further experiments, experiment definition was the most time-consuming, as it required experts from two countries, often not speaking the language of each other, precisely to understand experiment requirements and the limitations of each other's datasets. The informal nature of both data and experiment descriptions was a major stumbling block, and pointed towards the importance of formalisation efforts.

The time-cost of the process—three person-months of one-time bootstrapping effort followed by nine weeks for the experiment itself—needs to be considered in the context of the usual costs of setting up such experiments. As reported in section 3, for experiments *in Scotland alone*, the amount of time between receiving a request for experiment data and fulfilling it varies between six months and two years of elapsed time, with data preparation taking up to 10 months (elapsed) and one person-month of actual data processing work. Given the considerably higher complexity of cross-border integration, the cost of our semi-automated process is perfectly within the bounds of feasibility.

We conducted a scalability experiment on local data formalisation and cross-border mapping, which correspond to the fully automated part of the process (once the manual process definition has been carried out by data scientists). We used synthetic (yet realistic) data provided by NSS Scotland, that used the same schema as in the real-world experiment, but randomised values. The running times (Table 4) show a clearly regressing trend with respect to the number of records, which is due to the complexity of data transformations being linear with the number of *unique* labels. In real-world data, there tend to be frequent repetitions of a finite number of labels (e.g. in any given experiment, diseases will be limited to a relatively small subset of ICD codes). In our synthetic dataset, about 1,000 unique concepts needed to be dealt with. For this reason, the results of computations can be cached and reused efficiently. Note also that even for large datasets, running times remain negligible with respect to the magnitude of the full data preparation effort as reported in Table 3.

Table 3. Analysis of the duration of steps in terms of actual work time.

Step	Agents	Modality	Duration
knowledge bootstrapping	data scientists	semi-aut.	60 p/d
experiment definition	researchers, data scientists	manual	30 p/d (elapsed)
knowledge adaptation	researchers, data scientists	manual	5 p/d
local data formalisation	data scientists	semi-aut.	3 p/d
cross-border data mapping	data scientists	semi-aut.	2 p/d
validation of results	data scientists	manual	5 p/d

Table 4. Scalability of automated formalisation & cross-border mapping: running times on a synthetic dataset of Scottish death records.

Records	Time (s)	Records	Time (s)	Records	Time (s)
500	41	2,000	101	100,000	175
1,000	63	10,000	132	200,000	212

DOAC experiment results. The combined cohort size obtained from Scotland and Italy was 1,443. From a medical point of view, this amount of data turned out to be sufficient to cover only a part of the research initially planned. Meaningful results were obtained with respect to gender and age bias of patients, as well as with respect to relative risks of taking oral anticoagulants. On the other hand, the data size still was not statistically significant enough to compute comparative relative risks between DOACs and Warfarin. Finally, from the point of view of piloting the technology and methodology for cross-border data integration, the effort was successful as it confirmed the feasibility of our approach.

7 PERSPECTIVES

While the architecture is designed to scale with the addition of new data controllers, the experiment presented in this paper included only two participating data sources. The same approach, however, is being deployed in two ongoing projects that involve more jurisdictions and/or data controllers within the same jurisdiction. In the *InteropEHRate* EU-funded project[11], the cross-border interoperability of health records among countries of the European Union is solved through a FHIR-based data integration architecture. In another, Scotland-specific *Sprint Exemplar* project funded by the *UK Health Data Research Alliance*, data from multiple Scottish data controllers, 'regional Safe Havens', are harmonised through local integration, before cross-border integration takes place.

This paper and the underlying research were supported by EIT Digital, the University of Edinburgh, as well as the European Union's H2020 research and innovation programme under grant agreement No 826106, project *InteropEHRate*. We warmly thank Simone Bocca, Danish Cheema, David Leoni, Clifford Nangle, and Alessio Zamboni for their invaluable technical contributions.

[11] http://www.interopehrate.eu

REFERENCES

[1] Gábor Bella, Fiona McNeill, David Leoni, Francisco José Quesada Real, and Fausto Giunchiglia, 'Diversicon: Pluggable lexical domain knowledge', *Journal on Data Semantics*, **8**(4), 219–234, (2019).
[2] Gábor Bella, Fausto Giunchiglia, and Fiona McNeill, 'Language and Domain Aware Lightweight Ontology Matching', *Web Semantics: Science, Services and Agents on the World Wide Web*, **43**(1), (2017).
[3] Gábor Bella, Alessio Zamboni, and Fausto Giunchiglia, 'Domain-Based Sense Disambiguation in Multilingual Structured Data', in *The Diversity Workshop at the European Conference on Artificial Intelligence*, (2016).
[4] Olivier Bodenreider, 'The unified medical language system (UMLS): integrating biomedical terminology', *Nucleic acids research*, **32**(suppl_1), D267–D270, (2004).
[5] Olivier Bodenreider, 'Biomedical ontologies in action: role in knowledge management, data integration and decision support', *Yearbook of medical informatics*, **17**(01), 67–79, (2008).
[6] Subhashis Das and Fausto Giunchiglia, 'Geoetypes: Harmonizing diversity in geospatial data (short paper)', in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pp. 643–653. Springer, (2016).
[7] Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella, 'Understanding and exploiting language diversity.', in *IJCAI*, pp. 4009–4017, (2017).
[8] Fausto Giunchiglia, Biswanath Dutta, and Vincenzo Maltese, 'From knowledge organization to knowledge representation', *KO KNOWLEDGE ORGANIZATION*, **41**(1), 44–56, (2014).
[9] Fausto Giunchiglia, Vincenzo Maltese, and Biswanath Dutta, 'Domains and context: first steps towards managing diversity in knowledge', *Journal of web semantics*, **12**, 53–63, (2012).
[10] Joseph Goedert, 'New uses, clients for ibm watson health', *Health Data Management*, (2015).
[11] Shubham Gupta, Pedro Szekely, Craig A Knoblock, Aman Goel, Mohsen Taheriyan, and Maria Muslea, 'Karma: A system for mapping structured sources into the semantic web', in *Extended Semantic Web Conference*, pp. 430–434. Springer, (2012).
[12] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, 'A survey on deep learning in medical image analysis', *Medical image analysis*, **42**, 60–88, (2017).
[13] George A. Miller, 'WordNet: A Lexical Database for English', *Commun. ACM*, **38**(11), 39–41, (November 1995).
[14] David Pérez-Rey, Victor Maojo, Miguel García-Remesal, Raúl Alonso-Calvo, Holger Billhardt, Fernando Martín-Sánchez, and A Sousa, 'Ontofusion: Ontology-based integration of genomic and clinical databases', *Computers in biology and medicine*, **36**(7-8), 712–730, (2006).
[15] Fabian Prasser, Oliver Kohlbacher, Ulrich Mansmann, Bernhard Bauer, and Klaus A Kuhn, 'Data integration for future medicine (difuture)', *Methods of information in medicine*, **57**(S 01), e57–e65, (2018).
[16] Giuseppe Roberto, Ingrid Leal, Naveed Sattar, A Katrina Loomis, Paul Avillach, Peter Egger, Rients Van Wijngaarden, David Ansell, Sulev Reisberg, Mari-Liis Tammesoo, et al., 'Identifying cases of type 2 diabetes in heterogeneous data sources: strategy from the emif project', *PloS one*, **11**(8), (2016).
[17] David Robertson, Fausto Giunchiglia, Stephen Pavis, Ettore Turra, Gabor Bella, Elizabeth Elliot, Andrew Morris, Malcolm Atkinson, Gordon McAllister, Areti Manataki, et al., 'Healthcare data safe havens: towards a logical architecture and experiment automation', *The Journal of Engineering*, **2016**(11), 431–440, (2016).
[18] ST Rosenbloom, RJ Carroll, JL Warner, ME Matheny, and JC Denny, 'Representing knowledge consistently across health systems', *Yearbook of medical informatics*, **26**(01), 139–147, (2017).
[19] Casey Ross and Ike Swetlitz, 'Ibm pitched its watson supercomputer as a revolution in cancer care. it's nowhere close', *STAT Investigation*, (2017).
[20] Elisabeth Scheufele, Dina Aronzon, Robert Coopersmith, Michael T McDuffie, Manish Kapoor, Christopher A Uhrich, Jean E Avitabile, Jinlei Liu, Dan Housman, and Matvey B Palchuk, 'transmart: an open source knowledge management and high content data analytics platform', *AMIA Summits on Translational Science Proceedings*, **2014**, 96, (2014).