Effective queries for mega-analysis in cognitive neuroscience

Anna Ravenschlag[1], Monique Denissen[1], Bianca Löhnert[1,2], Mateusz Pawlik[1,*], Nicole Himmelstoß[1] and Florian Hutzler[1]

[1] *Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Hellbrunnerstraße 34, 5020 Salzburg, Austria*

[2] *Paris-Lodron-University of Salzburg, Department of Computer Sciences, Jakob-Haringer-Straße 2, 5020 Salzburg, Austria*

Abstract

Functional neuroimaging investigates the neural correlates of performing cognitive tasks. The empirical evidence in this field is constantly growing and gave rise to methods for assessment and integration of the results across different studies. A promising and suitable technique is the so-called mega-analysis. Performing mega-analysis is, however, challenging. It is a multi-step process which connects a researcher's implicit reasoning about information processing in the brain with complex analysis of heterogenous data. Although the process of mega-analysis is well understood, it comprises many concepts and queries that lack a formal definition. Therefore, it is difficult to choose a suitable data model, design a data schema, and implement the relevant queries. A prerequisite for a successful mega-analysis is a set of studies conforming to a carefully defined experimental setting. Finding such datasets is, however, a laborious and error-prone task of keyword-based literature search. To aid understanding of the underlying issues, we propose a conceptual model of mega-analysis. The model integrates a researcher's implicit knowledge with a systematic definition of relevant data. The nature of the data suggests a graph data model for effectively querying datasets. Consequently, we define a knowledge graph integrating the data associated with experimental setting, formally define the queries over the knowledge graph, and showcase their implementation in a graph database.

Keywords

mega-analysis, conceptual modeling, knowledge graph, graph queries, graph database, cognitive neuroscience

1. Introduction

Mega-analysis in functional neuroimaging. Functional neuroimaging is a method employed in cognitive neuroscience to investigate the relationships between cognitive processes and their neural correlates in the brain. Although cognitive processes cannot be directly observed, their neural correlates can be measured using neuroimaging technologies such as functional magnetic resonance imaging (fMRI). Therefore, the presence of a cognitive process is often inferred by contrasting brain activation under varying and carefully designed experimental settings.

The number of neuroimaging studies and the resulting datasets have grown exponentially in recent years [1]. The multitude of data sources and varying experimental settings make it crucial to assess and integrate findings across studies. This is commonly done by performing **meta**-analysis, i.e. an analysis of aggregated *results* reported in literature to assess how well they converge across studies. A downside of this approach is that reported results are a reduction of the orginally collected data. Neuroimaging acquires complex spatio-temporal data which contains more information than can be summarized in literature. Additionally, bias in literature is easily carried over into a meta-analysis. An analysis of aggregated *original data*, called **mega**-analysis, has been proposed as an alternative that has the potential to achieve increased statistical power and reliability compared to meta-analysis [2, 3]. A successful mega-analysis requires a set of datasets sharing homogenous experimental settings. To meet this requirement several efforts have emerged. The Brain Image Data Structure (BIDS) [4] is a common format standard for organizing data resulting from a neuroscientific experiment. Hierarchical Event Descriptors (HED) [5, 6] provides a taxonomy for describing details of the experimental setting. Online data repositories, like OpenNeuro [7, 8], al-

Published in the Workshop Proceedings of the EDBT/ICDT 2023 Joint Conference (March 28-March 31, 2023, Ioannina, Greece)

[*] Corresponding author.

annanatali.ravenschlag@plus.ac.at (A. Ravenschlag); monique.denissen@plus.ac.at (M. Denissen); bianca.loehnert@plus.ac.at (B. Löhnert); mateusz.pawlik@plus.ac.at (M. Pawlik); nicolealexandra.himmelstoss@plus.ac.at (N. Himmelstoß); florian.hutzler@plus.ac.at (F. Hutzler)

0000-0003-0272-1090 (A. Ravenschlag); 0000-0002-7120-1075 (M. Denissen); 0000-0003-0543-6332 (M. Pawlik); 0000-0003-0994-3649 (N. Himmelstoß); 0000-0001-8195-4911 (F. Hutzler)

low researchers to share their data with each other. OpenNeuro requires data to be stored according to BIDS, and currently houses more than 800 datasets. HED annotations are part of the BIDS specification. Despite these efforts, identifying the studies relevant to a mega-analysis remains a challenging task.

Mega-analysis workflow. A mega-analysis typically investigates a particular cognitive process or functionality of a brain region. The respective research question usually involves contrasting two or more carefully defined *experimental setting conditions* [9]. The criteria for a desired condition [10] can be arbitrarily complex. We exemplify three aspects of experimental settings: specifics of an activity the participants were tasked to perform, demographics and other characteristics of participants, and data acquisition parameters specific to a measuring device (see Example 1.1). Note that for a particular mega-analysis additional properties may also be of importance.

Ideally, the experimental settings should be queried directly in the datasets. This, however, is currently not possible due to poor data availability and annotation. Therefore, to find relevant datasets, a researcher starts with a keyword-based literature search followed by data requests. Keyword search is error-prone because it strongly depends on the presence of keywords in an article text and is biased by the choice of keywords [11]. We argue that introducing more frameworks similar to the one presented here may incentivize a higher availability of data and their more accurate annotation.

Once the qualifying datasets are collected, they undergo a signal analysis. The existing techniques vary, and their choice depends on the research question. The data acquired by fMRI can be interpreted as the changes in the intensity of brain activation at a specific brain location, called *voxel*, over time. A common analysis method models an expected brain activation on the basis of the specified experimental setting conditions, which in turn is compared to the recorded signal in each voxel. This analysis allows to explore whether the variation of conditions can explain an intensified brain activation at any location in the brain. The results for each participant of a study are aggregated, compared between different participant groups, and extrapolated to the whole population.

To better illustrate the concepts throughout this paper, we introduce the following example.

Example 1.1. *Running example.* A cognitive neuroscientist investigates the following research question: *Is there a difference in brain activation between processing the sex of male and female faces and does this differ between men and women?* The desired experimental settings for qualifying datasets are: *While recording a fMRI signal (acquisition parameters), men and women (participant demographics) were asked to identify the sex of faces presented in a series of images (activity details).* In this paper, we focus on the details of the activity the participants were tasked to perform.

Problem statement. The goal of this paper is a querying framework for mega-analysis in functional neuroimaging. We focus on solving the following problem: From a collection of datasets return those comprising a given experimental setting condition.

Challenges of querying experimental settings. Querying experimental settings effectively and transparently is essential, not only in the context of mega-analysis, but also to related work search or methodologies like reverse inference [10]. Unfortunately, it currently faces three main challenges.

1. The experimental setting is not systematically defined. A fundamental part of verifying the relevance of a dataset is matching its experimental setting to the desired conditions for the mega-analysis. Despite this, experimental settings lack a formal definition, which makes comparing studies difficult. The annotations provided in the data use arbitrary terminology making it impossible to infer the required details. More information can be found only in the publications describing the studies. At the same time, researchers focus their annotations on the study they conducted and not on possible future analysis scenarios.

2. The data is heterogenous in its format. The commonly applied BIDS format [4] specifies how datasets should be organized in a file system, including directory structure, file names and their formats. However, the relevant information is spread among tabular (TSV) and structured (JSON) files and has only a partial schema, which can be arbitrarily extended by user-defined columns and keys. Querying information in a file system is complicated and may be inefficient. Furthermore, it is hard to choose a suitable data model without fully understanding the data.

3. Querying experimental settings is limited to keyword search. The experimental settings are either narratively described in the corresponding publications or they are poorly annotated with arbitrary labels in the data. The custom and unstructured annotations vary between datasets and researchers, limiting the queries to imprecise keyword matching. Curated taxonomies like HED [5, 6] solve the problem of inconsistent terminology. The HED taxonomy consist of terms designed to describe experimental events on a level relevant to the study of human action, perception and cognition. HED terms can be grouped together to form a description of a particular aspect of an event. The experiment events can be annotated with one or more of these groups. However, the result is a collection of complex string labels, which cannot be queried directly.

Contributions. To address the challenges of building an effective querying framework for mega-analysis, this paper makes the following contributions.

- We propose a novel comprehensive model for mega-analysis that integrates the complex data with the researcher's reasoning. Our model captures the essential concepts down to the level of data types. This helps to determine the data model and a possible solution to the problem statement query.
- We propose a novel graph definition of experimental settings which is suitable for both querying conditions and data annotation. Such a representation allows us to simplify finding qualifying datasets with an elementary subgraph query.
- We showcase our solution by building a knowledge graph for the running example in Neo4j and implementing the queries with Cypher `MATCH` statements.

Thus, these contributions demonstrate a proof-of-concept for ontologically conceptualizing mega-analysis, translating the elements critical for dataset queries into a graph representation, and implementing a queriable prototype using Neo4j. The targeted user group for subsequent large-scale implementations are cognitive neuroscientists in need of effective and easy-to-use solutions for finding datasets suitable for mega-analysis.

2. Modeling mega-analysis

We present a conceptual model of a mega-analysis use case in Figure 1. It serves the purpose of clarifying relevant concepts in mega-analysis, their relations, and value spaces to derive an appropriate graph representation of the elements critical for querying datasets (see Section 3). For building the conceptual model, we ontologically analyzed the process of mega-analysis in cognitive neuroscience by identifying its constituent entities and how they relate to each other. Importantly, the resulting model represents a conceptualization that is tailored towards the requirements of our use case, i.e. it abstracts from reality to focus on the elements that are essential for our queries. As basis for the ontological analysis, we leveraged the environment of theories and tools provided by the Unified Foundational Ontology (UFO) which formally defines fundamental conceptual modeling notions such as entity and relationship types [12]. For the purpose of our current contribution, we employed the core categories of UFO (UFO-A) which describe endurant types such as objects, taxonomic relations, and associations. In future iterations we may further specify the conceptualization of mega-analysis by incorporating more recent developments on perdurant types (UFO-B) or intentional and social entities (UFO-C). The model was implemented using the ontology-driven conceptual modeling language OntoUML [13] which is based on the Unified Modeling Language. Compared to traditional conceptual modeling languages, OntoUML offers two main advantages for modeling our use case scenario: First, it allows for conceptual clarification by reflecting the ontological distinctions put forward by UFO. In traditional modeling languages such as OWL or UML, the ontologically distinctive types of entities and relations (which are made explicit in OntoUML) are collapsed to one single type of entity (e.g. class) and relation (e.g. association). Consequently, OntoUML provides means to differentiate various object and relationship types that reflect real-world semantics. Second, OntoUML introduces constraints which exclude the creation of models that would break the axiomatization of UFO, thus allowing researchers to explicate their domain specific knowledge within syntactically and semantically valid models.

The entities in our model are conceptualized as *kinds*, i.e. basic types of objects that exist in the real world and provide a uniform principle of identity for their instances. To represent the intrinsic properties of kinds, we employ *quality* types and *subkinds* of quality types, i.e. functions that take elements in the extension of an object type and map them to a respective quality structure. These quality structures form either one-dimensional (quality dimension) or multi-dimensional (quality domain) conceptual spaces. In OntoUML, quality structures are represented as *datatypes* that organize the possible values which can be attributed to the respective quality types. The relationships between kinds are modeled as *material relations* that are existentially dependent on both their bearer and an external entity, i.e. they link two kinds by establishing a

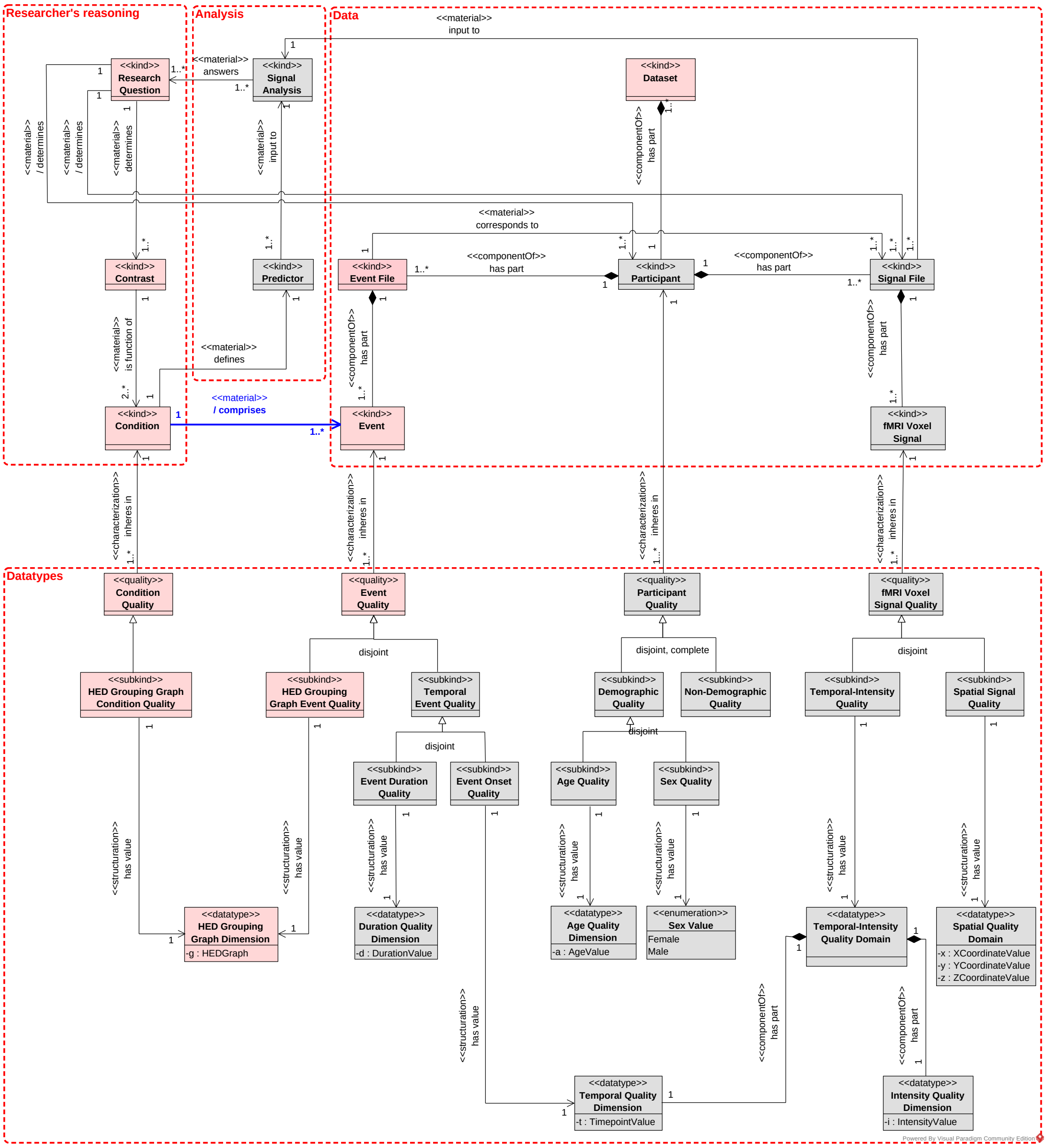
Researcher's reasoning
Analysis
Data
Datatypes
<<kind>> Research Question
<<kind>> Signal Analysis
<<kind>> Dataset
<<kind>> Contrast
<<kind>> Predictor
<<kind>> Event File
<<kind>> Participant
<<kind>> Signal File
<<kind>> Condition
<<kind>> Event
<<kind>> fMRI Voxel Signal
<<material>> answers
<<material>> input to
<<material>> / determines
<<material>> determines
<<material>> corresponds to
<<componentOf>> has part
<<material>> is function of
<<material>> defines
<<material>> / comprises
<<characterization>> inheres in
<<quality>> Condition Quality
<<quality>> Event Quality
<<quality>> Participant Quality
<<quality>> fMRI Voxel Signal Quality
disjoint
disjoint, complete
<<subkind>> HED Grouping Graph Condition Quality
<<subkind>> HED Grouping Graph Event Quality
<<subkind>> Temporal Event Quality
<<subkind>> Demographic Quality
<<subkind>> Non-Demographic Quality
<<subkind>> Temporal-Intensity Quality
<<subkind>> Spatial Signal Quality
<<subkind>> Event Duration Quality
<<subkind>> Event Onset Quality
<<subkind>> Age Quality
<<subkind>> Sex Quality
<<structuration>> has value
<<datatype>> HED Grouping Graph Dimension
-g : HEDGraph
<<datatype>> Duration Quality Dimension
-d : DurationValue
<<datatype>> Age Quality Dimension
-a : AgeValue
<<enumeration>> Sex Value
Female
Male
<<datatype>> Temporal-Intensity Quality Domain
<<datatype>> Spatial Quality Domain
-x : XCoordinateValue
-y : YCoordinateValue
-z : ZCoordinateValue
<<datatype>> Temporal Quality Dimension
-t : TimepointValue
<<datatype>> Intensity Quality Dimension
-i : IntensityValue
1
1..*
2..*
Powered By Visual Paradigm Community Edition

Figure 1: Conceptual model of mega-analysis in OntoUML composed of researcher's reasoning (top left), analysis (top middle), data (top right), and datatypes (bottom).

semantically meaningful connection between them. Qualities inhere in their bearer entities via *characterization relations*, i.e. they represent the features intrinsic to the object type they existentially depend on. Qualities, in turn, are structured via *structuration relations*, connecting them to the datatypes that define the space of possible values a quality can take.

As outlined in Figure 1, performing a mega-analysis necessitates to connect i) the researcher's reasoning (top left) with ii.) the planned analysis (top middle), iii.) the datasets that qualify for this analysis by means of a particular experimental setting (top right), and iv.) the different data types comprised in the datasets (bottom). Since events are the essential building blocks of the experimental setting, our current work focuses exclusively on the connection between the researcher's reasoning and event data via a common, queriable datatype (colored entities). For a standardized description of event data, we employ the HED taxonomy [5, 6]. Whether a dataset qualifies for a mega-analysis depends on whether it matches the desired experimental setting. In terms of events, the experimental setting of a dataset is captured in the HED annotations. In order to query the data for our mega-analysis we also define our conditions in HED annotations, thus establishing a common data format between the conditions derived from the researcher's reasoning and HED annotated event data.

Example 2.1. The *research question* defined in our running example (Example 1.1) determines one or more *contrasts* that would qualify to assess this question, e.g. *male faces presented with an instruction to identify sex versus female faces presented with an instruction to identify sex.* In neuroimaging research, such a contrast is commonly defined as a function of two or more experimental setting *conditions*, in this case *male face identification* and *female face identification.* We can express these conditions in HED annotations strings using ((Face, Human-agent, Female), (Task, (Discriminate, Sex))) and ((Face, Human-agent, Male), (Task, (Discriminate, Sex))). These *condition qualities* can be represented within the space of a *HED grouping graph dimension* (see Section 3 for a formal definition). Experimental setting conditions, in turn, comprise the specific *events* that need to be stored in the *event files* of a *dataset* so that it qualifies for use in the researcher's mega-analysis. By projecting *event qualities* into the same value space of the HED grouping graph dimension, we guarantee a mutual data type between the events that are i.) present in a dataset and ii.) required by the desired conditions for the mega-analysis, thus enabling effective querying.

Although it is beyond the scope of the current paper, we incorporated the remaining parts of mega-analysis as greyed-out entities in the model for completeness. For example, qualities and datatypes associated with the participants of datasets can determine additional aspects of the desired experimental setting, e.g. with respect to a specific age range. Since performing a full neuroimaging mega-analysis involves a complex, multi-step analysis, our conceptual model also includes a representation of the acquired signal with its qualities and value spaces as these are pertinent to ultimately answer the research question. Note that the analysis is dependent on the same information that we use for querying. Finding the conditions in a dataset is thus not only relevant during the searching stage of a mega-analysis, but also for the analysis part, which we aim to address in future work.

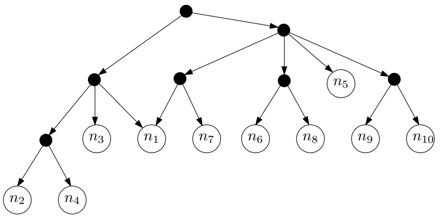

Figure 2: HED Grouping Graph $\mathcal{H}_1$

The following sections formally define how to computationally derive the link between relevant conditions for a mega-analysis, specified by the researcher's reasoning, and events in a dataset (blue relationship in Figure 1). Subsequently, we demonstrate an implementation of data and queries in a graph database.

3. Experimental setting conditions

Our conceptual model indicates that datasets qualifying for a mega-analysis can be found by matching the HED annotations of experimental setting conditions and events (blue relationship in Figure 1). HED annotations are stored in the data as long string values that are difficult to query. In fact, they can form arbitrarily nested groups of terms, i.e., a graph. A graph of a single HED annotation is called a *HED grouping graph* and the set of all terms from the HED taxonomy is depicted by *HED*.

Definition 3.1 (HED grouping graph)**.** A *HED grouping graph* $\mathcal{H}$ with nodes $N(\mathcal{H})$ and edges $E(\mathcal{H})$ is a directed, connected, acyclic graph with exactly one node without incoming edges, $R(\mathcal{H})$, called the *root.* Nodes without outgoing edges are called *leaves.* Each leaf node n has a label $L(n) \in \mathit{HED}$.

Example 3.1. Figure 2 shows a HED grouping graph $\mathcal{H}_1$ with the following leaf node labels:

$L(n_1) = \mathit{Face}$	$L(n_2) = \mathit{Rotated}$
$L(n_3) = \mathit{Male}$	$L(n_4) = \mathit{Downward}$
$L(n_5) = \mathit{Task}$	$L(n_6) = \mathit{Discriminate}$
$L(n_7) = \mathit{Detect}$	$L(n_8) = \mathit{Sex}$
$L(n_9) = \mathit{Press}$	$L(n_{10}) = \mathit{Push\text{-}button}$

Using the HED grouping graphs, we define the data concepts of our model (red entities in the top-right data section of Figure 1).

Definition 3.2 (Event, event file, dataset)**.** An *event* e is a triple of the form $(o(e), d(e), \mathcal{H}(e))$, where $o(e) \in \mathbb{R}$ is the onset (the timepoint when the event e started), $d(e) \in \mathbb{R}^+$ is the duration of e and $\mathcal{H}(e)$ is a HED grouping graph. An *event file* F is a set of events and a *dataset* D is a set of event files.

To facilitate returning datasets as the results of the queries, we introduce a *dataset graph* $\mathcal{D}(D)$ which is composed of:

- a *dataset node* n_D,
- one *event file node* n_F for every event file $F \in D$,
- all HED grouping graphs in the event files of D,
- edges between the dataset node and all event file nodes,
- edges between the event file nodes and the root nodes of all respective HED grouping graphs.

Definition 3.3 (Dataset graph)**.** Let $H(F) = \bigcup_{e \in F} \mathcal{H}(e)$ be the union of all HED grouping graphs from an event file F in a dataset D. The *dataset graph* $\mathcal{D}(D)$ is the union of all HED grouping graphs $\bigcup_{F \in D} H(F)$ with additional nodes $n_D \cup \{n_F \mid F \in D\}$ and edges $\{(n_D, n_F) \mid F \in D\} \cup \{(n_F, R(\mathcal{H})) \mid F \in D \wedge \mathcal{H} \in H(F)\}$.

Example 3.2. Consider a dataset D with a single event file $F = \{(1, 1.5, \mathcal{H}_1)\}$, where $\mathcal{H}_1$ is the HED grouping graph in Figure 2. Then, the dataset graph $\mathcal{D}(D)$ has nodes $\{n_D, n_F\} \cup N(\mathcal{H}_1)$ and edges $\{(n_D, n_F), (n_F, R(\mathcal{H}_1))\} \cup E(\mathcal{H}_1)$.

To perform a mega-analysis, the researcher defines a set of experimental setting conditions. For a dataset to qualify for the mega-analysis, it must contain at least one event for each of the specified conditions. We define an experimental setting condition as a HED grouping graph. Thus, the matching between events and conditions can be resolved over HED grouping graphs.

A *condition query* returns all datasets from a collection $\mathbf{S}$ of dataset graphs containing a subgraph which matches exactly the HED grouping graph of the specified condition. In this paper we are interested in exact subgraph matches. In the future, we plan to make the condition query more flexible.

Definition 3.4 (Condition query)**.** Given a set $\mathbf{S}$ of dataset graphs and a condition $\mathcal{H}$, the *condition query* $Q(\mathbf{S}, \mathcal{H})$ is defined as follows:

$$Q(\mathbf{S}, \mathcal{H})=\{D \mid \mathcal{D}(D) \in \mathbf{S} \wedge \mathcal{H} \text{ is subgraph of } \mathcal{D}(D)\}$$

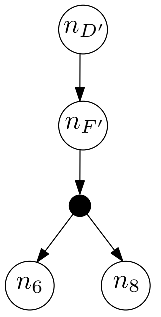

Figure 3: $\mathcal{D}(D')$

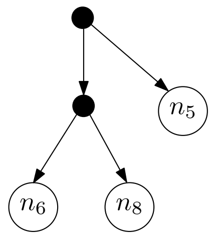

Figure 4: $\mathcal{H}_c$

Example 3.3. In Example 2.1 we identified HED annotations for the conditions that are relevant to the research question of our running example. For demonstration purposes we consider a simplified version of one of these conditions, namely (Task, (Discriminate, Sex)). In other words, the researcher is interested in all datasets from a collection $\mathbf{S}$ that contain this particular condition. The resulting datasets compose the input to the mega-analysis of interest. The HED grouping graph $\mathcal{H}_c$ of the above condition is shown in Figure 4 and let $\mathbf{S}$ be the set $\{\mathcal{D}(D), \mathcal{D}(D')\}$, where $\mathcal{D}(D)$ is the dataset graph from Example 3.2 and $\mathcal{D}(D')$ is the dataset graph in Figure 3. The leaf node labels of $\mathcal{D}(D)$, $\mathcal{D}(D')$ and $\mathcal{H}_c$ are as in Example 3.1. Then the condition query $Q(\mathbf{S}, \mathcal{H}_c) = \{D\}$ returns only the dataset D, since $\mathcal{H}_c$ is a subgraph of $\mathcal{D}(D)$, but not of $\mathcal{D}(D')$.

4. Implementation in Neo4j

In this section we demonstrate a prototype implementation which enables cognitive neuroscientists with effective querying for datasets relevant to the desired mega-analysis. For this purpose, we manually annotated 35 datasets from Openneuro[7] and acquired at the Centre for Cognitive Neuroscience, University of Salzburg. We implemented a prototype data graph and the queries in *Neo4j* [14] graph database management system. The datasets are loaded into Neo4j as dataset graphs (cf. Section 3) using our custom indexer [15]. The resulting data graph in Neo4j has around 720k nodes, 1.7M relationships and 254 unique HED grouping graphs.

Cypher queries. The querying language of the Neo4j database system is *Cypher* [16]. In order to execute a condition query (cf. Definition 3.4) on the graph in Neo4j, we need to translate it into Cypher. Given a condition HED grouping graph

$\mathcal{H}_c$, Algorithm 1 generates the respective Cypher query. For the pattern matching that is necessary to return the required datasets Algorithm 1 uses `MATCH` clause of Cypher.

```
   Input  : Condition HED grouping graph $\mathcal{H}_c$
   Output : Cypher query $C(\mathcal{H}_c)$
1  begin
2      $C(\mathcal{H}_c)$ ={MATCH (D:Dataset)}
3      foreach leaf $n_l \in N(\mathcal{H}_c)$ do
4          $C(\mathcal{H}_c) = C(\mathcal{H}_c) \cup \{(n_l : L(n_l))\}$
5      end
6      foreach $(n, n') \in E(\mathcal{H}_c)$ do
7          $C(\mathcal{H}_c) = C(\mathcal{H}_c) \cup \{(n) \rightarrow (n')\}$
8      end
9      $C(\mathcal{H}_c) = C(\mathcal{H}_c) \cup$ {(D)-[*]->($R(\mathcal{H}_c)$)}
10     $C(\mathcal{H}_c) = C(\mathcal{H}_c) \cup$ {RETURN D }
11     return $C(\mathcal{H}_c)$
12 end
```

Algorithm 1: Generate Cypher query $C(\mathcal{H})$ for the condition $\mathcal{H}_c$.

Example 4.1. Consider the condition HED grouping graph $\mathcal{H}_c$ of Example 3.3 and a set of dataset graphs that resides in the Neo4j database. To enable the execution of the query in Neo4j we apply Algorithm 1 and translate it to Cypher, the respective query $C(\mathcal{H}_c)$ looks as follows:

$$
\begin{aligned}
C(\mathcal{H}_c) = \{\texttt{MATCH (D:Dataset)}, \quad & (n_5 : \texttt{Task}),\\
(n_6 : \texttt{Discriminate}), \quad & (n_8 : \texttt{Sex}),\\
(r) \rightarrow (n_5), \quad & (r) \rightarrow (n_{68}),\\
(n_{68}) \rightarrow (n_6), \quad & (n_{68}) \rightarrow (n_8),\\
\texttt{(D)-[*]->}(r) & \\
\texttt{RETURN D}\} &
\end{aligned}
$$

Note that the node identifier r, n_5, n_6, n_8 and n_{68} emerge from line 7 of Algorithm 1 and are arbitrary Cypher variable names. Executing this query on our data graph in Neo4j returns three datasets. The researcher can then perform mega-analysis procedures with these three datasets and interpret the results of the analysis to answer the research question. This demonstrates that our approach allows the researcher to find all datasets that are relevant to the research question. We achieve this based on HED annotations at the level of individual events, in contrast to other approaches (cf. Section 5), where researchers must limit themselves to using keywords or predefined labels to find relevant datasets for mega-analysis. Another advantage is that researchers know the context of the query result, i.e. why a dataset is returned, and thus can verify whether the resulting datasets are appropriate without reading the related publications as well as modify the condition query when necessary.

5. Related work

Assessment and integration of the results across different studies in the field of cognitive neuroscience have, until now, mainly relied on aggregated results of previously performed analyses. Several systems have been designed to store and query such results. Moreover, there have been efforts to improve data annotation along with the development of systems for data storage. We list a selection of particularly influential systems and summarize their features in Table 1. These efforts, however, only offer partial solutions with respect to our use case of effective data querying for mega-analysis. To enable a successful mega-analysis, we identify the following three aspects of system requirements.

Data: The original data acquired in a study must be available. In contrast, aggregated data derivatives resulting from analysis, e.g., so-called peak coordinates and statistical maps, are not sufficient for mega-analysis [2]. Although data availability is not the focus of this paper, it is a prerequisite for the mega-analysis use case. Our solutions can be applied to querying original data from various repositories.

Experimental setting: A suitable annotation schema uses a controlled taxonomy of terms and allows to describe all relevant aspects of an experimental setting, especially at the level of events. An example of such a taxonomy is HED [5, 6]. Arbitrary labels not only dramatically reduce the number of qualifying datasets that can be found by a query but may also cause false positives. The descriptions must be available on an event level, not just to ensure precision of the query, but also to execute the analysis once datasets have been identified. Datasets that lack this level of annotation cannot be processed for the purpose of a mega-analysis.

Queries: Ideally, a researcher can define custom experimental setting conditions in a systematic way and find all qualifying datasets. Querying arbitrary, non-standardized string labels requires guessing the right label and potentially results in missing relevant datasets. Moreover, querying aspects other than experimental setting conditions requires further investigation of the resulting datasets.

In Table 1 we show that none of the existing systems satisfies all requirements that we defined.

BrainMap, NeuroSynth, and NeuroVault focus on storing data derivatives. Brainmap and NeuroSynth

Table 1
Mega-analysis requirements addressed in existing systems

System	Data	Experimental setting	Queries
OpenNeuro [7, 8]	original	HED (few datasets only)	keywords
BrainMap [17, 18, 19]	derivatives only	BrainMap taxonomy	only existing labels
NeuroVault [20, 21]	derivatives only	arbitrary labels	keywords
NeuroSynth [22, 23]	derivatives only	keywords from publication texts	keywords
NeuroScout [24, 25]	original	ML classifier labels	ML classifier labels
PubMed [26]	publications	publication text	keywords

store so-called peak coordinates, which are a significant reduction from original data. NeuroVault stores statistical maps of the brain, which carry more information than peak coordinates, yet they still only represent derivations of the original data.

Data stored in Brainmap has been manually extracted and annotated from the literature. It consists of over 4000 scientific publications and 21000 contrast analyses. It uses a custom annotation schema [17, 19, 18]that only allows to describe experimental settings at the level of conditions, which is sufficient for peak coordinates data. The schema prevents querying arbitrary conditions which is necessary for specifying a desired mega-analysis. Additionally, the descriptions mix the standardized terms with free text annotations and querying is limited to choosing from a list of existing labels.

NeuroSynth data has been extracted and annotated from the literature using automated text analysis [22, 23]. It accumulates results in form of peak coordinates from over 13000 publications. Users can query the data with single terms or their sets organized into topics.

The data stored in NeuroVault is annotated manually with arbitrary labels. There is no schema and the annotations often contain abbreviations and study specific terms. Additionally, the statistical maps are stored as part of a collection, which has a general description. Both the description and label can be queried using keyword search.

OpenNeuro [7, 8] and NeuroScout [24, 25] focus on original data. OpenNeuro is a data repository that stores a wide variety of neuroimaging data. NeuroScout is a portal to a small number of curated datasets from OpenNeuro that share an experimental setting [24, 25]. OpenNeuro stores original data in BIDS format which allows for HED annotations [7, 8]. Unfortunately, the datasets are not curated and the majority does not include HED annotations. Moreover, the querying features of OpenNeuro are limited to keywords extracted from dataset descriptions.

In all the datasets stored by NeuroScout, the experimental setting is a form of a continuous narrative, e.g. a movie or an audio recording. Experimental settings are automatically annotated using various machine learning feature extraction techniques which also predefine the available querying terms. NeuroScout is a valuable resource, but the data is a limited sample of what is collected in the field of cognitive neuroscience.

Although PubMed is a repository of scientific publications, we decided to list it, as it is a frequently used tool for finding relevant publications for analysis across studies [26]. Keywords are queried in the publication texts and data can be obtained only by using information available in the publications.

To minimize the effort and maximize the amount of annotations, several of the systems listed in Table 1 automatically label the data. We briefly describe them and explain why they are insufficient for the use case of mega-analysis. NeuroSynth automatically extracts terms from scientific publication texts which are subsequently manually filtered for relevance in the field of neuroscience. Unfortunately, the publications primarily describe analyses, their results and how they contribute to the field. Information about the data, in particular the events that occurred during data acquisition, is often not provided, especially if it is not directly relevant for the analysis. Consequently, the adequacy of datasets returned by NeuroSynth is evaluated based on the original purpose of data collection, thus limiting the capacity of potential data-reuse. Note that none of the datasets that were returned by the query in Example 4.1 were originally collected to study processing of face sex, which is the focus of our example mega-analysis. Accordingly, this corroborates that our solution is capable of identifying datasets based on the events they comprise, hence effectively extrapolating the usability beyond the original intention.

NeuroScout uses machine learning classifiers to automatically annotate experimental settings in a complete recording of an experiment. However, this is currently only applicable to a limited sample of datasets where such recordings are available. More commonly, the software responsible for executing experiments provides only textual log files. A majority of the log files contain only abbreviations or numeric codes that cannot be understood without input from the original researchers. The BIDS specification and HED taxonomy allow for more comprehensive and structured annotation than can be achieved with automated approaches, and we hope to incentivize more widespread use of these tools.

6. Conclusion and future work

In this paper, we proposed a conceptual model for neuroimaging mega-analysis. We formally defined the parts of the model that are essential for finding qualifying datasets and implemented the resulting queries in Neo4j graph database. In a next step, we will demonstrate our solutions on a larger scale including more datasets and the remaining aspects of experimental settings. We plan to integrate the entire HED taxonomy into our knowledge graph and thus enrich the experimental setting annotations with otherwise implicit knowledge. An interesting extension of our work is to make the condition query more flexible by relaxing the subgraph constraint. Our long-term objective is to not only query the relevant datasets for a mega-analysis, but also to enable the execution of mega-analysis. To that extent, we will include the neuroimaging data, analysis workflows and their results into our framework, as highlighted in Section 2.

Acknowledgements. This work was supported by: the Austrian Federal Ministry of Education, Science and Research (BMBWF) under grant number 2920 (Austrian NeuroCloud); the Federal State of Salzburg under grant number 20102-F2101143-FPR (Digital Neuroscience Initiative); the Austrian Science Fund (FWF) under grant number W1233-B (Doctoral College "Imaging the Mind"). We thank Magdalena Ortiz and Fabio Richlan for their feedback and Ben Engler for annotating the datasets.

References

[1] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, F. J. Martinez, J. M. Gorriz, Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation, Information Fusion 64 (2020) 149–187. doi:10.1016/j.inffus.2020.07.006.

[2] S. G. Costafreda, Pooling FMRI data: Meta-analysis, mega-analysis and multi-center studies, Frontiers in Neuroinformatics 3 (2009) 33. doi:10.3389/neuro.11.033.2009.

[3] P. S. W. Boedhoe, M. W. Heymans, L. Schmaal, Y. Abe, P. Alonso, S. H. Ameis, A. Anticevic, P. D. Arnold, M. C. Batistuzzo, F. Benedetti, J. C. Beucke, I. Bollettini, A. Bose, S. Brem, A. Calvo, R. Calvo, Y. Cheng, K. I. K. Cho, V. Ciullo, S. Dallaspezia, D. Denys, J. D. Feusner, K. D. Fitzgerald, J.-P. Fouche, E. A. Fridgeirsson, P. Gruner, G. L. Hanna, D. P. Hibar, M. Q. Hoexter, H. Hu, C. Huyser, N. Jahanshad, A. James, N. Kathmann, C. Kaufmann, K. Koch, J. S. Kwon, L. Lazaro, C. Lochner, R. Marsh, I. Martínez-Zalacaín, D. Mataix-Cols, J. M. Menchón, L. Minuzzi, A. Morer, T. Nakamae, T. Nakao, J. C. Narayanaswamy, S. Nishida, E. L. Nurmi, J. O'Neill, J. Piacentini, F. Piras, F. Piras, Y. C. J. Reddy, T. J. Reess, Y. Sakai, J. R. Sato, H. B. Simpson, N. Soreni, C. Soriano-Mas, G. Spalletta, M. C. Stevens, P. R. Szeszko, D. F. Tolin, G. A. van Wingen, G. Venkatasubramanian, S. Walitza, Z. Wang, J.-Y. Yun, ENIGMA-OCD Working-Group, P. M. Thompson, D. J. Stein, O. A. van den Heuvel, J. W. R. Twisk, An Empirical Comparison of Meta- and Mega-Analysis With Data From the ENIGMA Obsessive-Compulsive Disorder Working Group, Frontiers in Neuroinformatics 12 (2018) 102. doi:10.3389/fninf.2018.00102.

[4] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, D. A. Handwerker, M. Hanke, D. Keator, X. Li, Z. Michael, C. Maumet, B. N. Nichols, T. E. Nichols, J. Pellman, J.-B. Poline, A. Rokem, G. Schaefer, V. Sochat, W. Triplett, J. A. Turner, G. Varoquaux, R. A. Poldrack, The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments, Scientific Data 3 (2016) 160044–160044. doi:10.1038/sdata.2016.44.

[5] K. Robbins, D. Truong, S. Appelhoff, A. Delorme, S. Makeig, Capturing the nature of events and event context using hierarchical event descriptors (HED), NeuroImage 245 (2021) 118766. doi:10.1016/j.neuroimage.2021.118766.

[6] K. Robbins, D. Truong, A. Jones, I. Callanan, S. Makeig, Building FAIR functionality: Annotating events in time series data using hierarchical event descriptors (HED), Neuroin-

formatics 20 (2022) 463–481. doi:10.1007/s12021-021-09537-4.

[7] Openneuro, 2023. URL: https://openneuro.org/.

[8] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, R. Blair, Y. O. Halchenko, E. Miller, N. Hardcastle, J. Wexler, O. Esteban, M. Goncavles, A. Jwa, R. Poldrack, The OpenNeuro resource for sharing of neuroscience data, eLife 10 (2021) e71774. doi:10.7554/eLife.71774.

[9] V. I. Müller, E. C. Cieslik, A. R. Laird, P. T. Fox, J. Radua, D. Mataix-Cols, C. R. Tench, T. Yarkoni, T. E. Nichols, P. E. Turkeltaub, T. D. Wager, S. B. Eickhoff, Ten simple rules for neuroimaging meta-analysis, Neuroscience and Biobehavioral Reviews 84 (2018) 151–161. doi:10.1016/j.neubiorev.2017.11.012.

[10] F. Hutzler, Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data, NeuroImage 84 (2014) 1061–1069. doi:10.1016/j.neuroimage.2012.12.075.

[11] M. Tahmasian, A. A. Sepehry, F. Samea, T. Khodadadifar, Z. Soltaninejad, N. Javaheripour, H. Khazaie, M. Zarei, S. B. Eickhoff, C. R. Eickhoff, Practical recommendations to conduct a neuroimaging meta-analysis for neuropsychiatric disorders, Human Brain Mapping 40 (2019) 5142–5154. doi:10.1002/hbm.24746.

[12] G. Guizzardi, A. B. Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. P. Sales, UFO: unified foundational ontology, Appl. Ontology 17 (2022) 167–210. doi:10.3233/AO-210256.

[13] G. Guizzardi, C. M. Fonseca, A. B. Benevides, J. P. A. Almeida, D. Porello, T. P. Sales, Endurant types in ontology-driven conceptual modeling: Towards ontouml 2.0, in: Conceptual Modeling - 37th International Conference, ER 2018, Xi'an, China, October 22-25, 2018, Proceedings, volume 11157 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 136–150. doi:10.1007/978-3-030-00847-5_12.

[14] Neo4j, 2023. URL: https://neo4j.com/.

[15] BIDS Indexer, 2023. URL: https://gitlab.com/ccns/neurocog/neurodataops/dni/bids-indexer.

[16] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, A. Taylor, Cypher: An evolving query language for property graphs, in: Proceedings of the 2018 International Conference on Management of Data, 2018, pp. 1433–1445. doi:10.1145/3183713.3190657.

[17] Brainmap, 2023. URL: http://brainmap.org/.

[18] P. T. Fox, J. L. Lancaster, Mapping context and content: The BrainMap model, Nature Reviews Neuroscience 3 (2002) 319–321. doi:10.1038/nrn789.

[19] P. T. Fox, A. R. Laird, S. P. Fox, P. M. Fox, A. M. Uecker, M. Crank, S. F. Koenig, J. L. Lancaster, Brainmap taxonomy of experimental design: description and evaluation, Human Brain Mapping 25 (2005) 185–98. doi:10.1002/hbm.20141.

[20] Neurovault, 2023. URL: https://neurovault.org/.

[21] K. J. Gorgolewski, G. Varoquaux, G. Rivera, Y. Schwarz, S. S. Ghosh, C. Maumet, V. V. Sochat, T. E. Nichols, R. A. Poldrack, J.-B. Poline, T. Yarkoni, D. S. Margulies, NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain, Frontiers in Neuroinformatics 9 (2015).

[22] Neurosynth, 2023. URL: https://neurosynth.org/.

[23] T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, T. D. Wager, Large-scale automated synthesis of human functional neuroimaging data, Nature Methods 8 (2011) 665–670. doi:10.1038/nmeth.1635.

[24] Neuroscout, 2023. URL: https://neuroscout.org/.

[25] A. de la Vega, R. Rocca, R. W. Blair, C. J. Markiewicz, J. Mentch, J. D. Kent, P. Herholz, S. S. Ghosh, R. A. Poldrack, T. Yarkoni, Neuroscout, a unified platform for generalizable and reproducible fMRI research, 2022. doi:10.1101/2022.04.05.487222.

[26] Pubmed, 2023. URL: https://pubmed.ncbi.nlm.nih.gov/.