AI Open 5 (2024) 55–69

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

Full length article

An ecosystem for personal knowledge graphs: A survey and research roadmap

Martin G. Skjæveland *, Krisztian Balog, Nolwenn Bernard, Weronika Łajewska, Trond Linjordet

University of Stavanger, Norway

ARTICLE INFO

Keywords:
Personal knowledge graphs
Personal data management

ABSTRACT

This paper presents an ecosystem for personal knowledge graphs (PKGs), commonly defined as resources of structured information about entities related to an individual, their attributes, and the relations between them. PKGs are a key enabler of secure and sophisticated personal data management and personalized services. However, there are challenges that need to be addressed before PKGs can achieve widespread adoption. One of the fundamental challenges is the very definition of what constitutes a PKG, as there are multiple interpretations of the term. We propose our own definition of a PKG, emphasizing the aspects of (1) data ownership by a single individual and (2) the delivery of personalized services as the primary purpose. We further argue that a holistic view of PKGs is needed to unlock their full potential, and propose a unified framework for PKGs, where the PKG is a part of a larger ecosystem with clear interfaces towards data services and data sources. A comprehensive survey and synthesis of existing work is conducted, with a mapping of the surveyed work into the proposed unified ecosystem. Finally, we identify open challenges and research opportunities for the ecosystem as a whole, as well as for the specific aspects of PKGs, which include population, representation and management, and utilization.

1. Introduction

The concept of a *personal knowledge graph* (PKG) broadly refers to "a resource of structured information about entities personally related to its user, their attributes and the relations between them" (Balog and Kenter, 2019). There are clear advantages to using PKGs for personal data management and storage as well as for supporting personalized services, such as search and recommendation: the user remains in control of their data and can decide what type of access to grant to which service on what part of their PKG.

While the overall vision is appealing, and there is a growing body of work on PKGs (Tiwari et al., 2023), several challenges remain before PKGs can deliver on their promise and enjoy more widespread adoption. One of the remaining open fundamental questions is the very definition of a PKG. There appear to be multiple interpretations of what constitutes a PKG, and often the term is only implicitly defined. Often, the concept is used in the sense of *personalized knowledge graphs* (or *personal interest graph*), i.e., a subset of an existing knowledge graph that characterizes the interests of a given individual. This interpretation leaves out the most essential feature of a *personal knowledge graph*: having the individual in control of their data (to decide who and what services get read or write access). As we will illustrate with two specific scenarios below, this distinction is critical for unlocking the full potential of PKGs. While Solid (Social Linked Data) is a prominent Semantic Web initiative promoting personal data storage in personal online data stores (PODs) (Sambra et al., 2016), its focus primarily lies on data ownership and management. In comparison, our work takes a more holistic approach to the PKG problem space, encompassing several areas where Solid currently falls short, such as seamless integration of various data sources and user-friendly administration.

The main contributions of this work are threefold. First, we present a unifying framework for PKGs, emphasizing that they need to form part of a larger ecosystem and interact with other services and data sources to reach their full potential. We argue that such a holistic view is needed in order to understand what the key features and requirements of PKGs are and how different research areas can contribute to addressing these. This holistic treatment of PKGs is perhaps the most important difference that distinguishes ours from previous work, which tends to focus on a single aspect or a narrow use of PKGs in a given application context. Second, we survey and synthesize existing work using our unified framework. Specifically, we organize our discussion around three main aspects of PKGs: (1) population, (2) representation and management, and (3) utilization. Third, we identify a set of open

* Corresponding author.
E-mail address: martin.g.skjeveland@uis.no (M.G. Skjæveland).

https://doi.org/10.1016/j.aiopen.2024.01.003
Received 18 April 2023; Accepted 29 January 2024
Available online 1 February 2024

challenges and outline research opportunities, both for the ecosystem as a whole and specific to each of the three main aspects listed above.

1.1. Motivating scenarios

To illustrate the use of PKGs, we present two motivating scenarios that will be used as running examples throughout the paper. The proposed scenarios differ significantly in their complexity and challenges involved. The first scenario addresses a straightforward utilization of a PKG for a personalized recommender system that exemplifies how information integrated from varying data sources may be leveraged for providing highly contextualized services. Whereas, the second scenario presents a more complex usage of a PKG for managing and sharing health information that involves additional challenges, such as handling inconsistencies between different data sources and dealing with sensitive data.

Example 1 (*Personal Trainer Assistant*)**.** Representing a family of applications around personalized recommender systems, a *personal trainer assistant* service can suggest a training plan using personal information regarding the physical condition, diet, and previous injuries of the person that are stored in a PKG. The training recommendations can be synchronized with the user's calendar and include different forms of activities depending on individual preferences. The PKG may integrate with external sources of personal information, for example, Facebook or YouTube, to access information about attended sport events or subscribed workout channels. External sources of public data such as Wikidata may also be integrated with the PKG to get more in-depth information about diseases, dietary requirements, or physical injuries. The personal trainer application is not limited to providing personalized recommendations in response to explicitly expressed information needs, but may include proactive suggestions related to other integrated services in response to user's activity. For example, it might invite the user to a suitable Strava[1] challenge if it becomes available at a nearby location.

Example 2 (*Sharing Health Information*)**.** In addition to forming a basis for personalized recommendations, a PKG may also constitute a tool for managing and sharing personal information. For example, in the case of a complex disease, a person may be a patient or client to a variety of different health service providers. In such a situation, it is typically a challenge for different services to stay updated on the current status of the person's disease and treatment (e.g., the current medication regimen and relevant medical facts in family history). Current clinical practice relies on documenting many important facts in prose text journals, which in turn are often too voluminous and dense for health personnel to exhaustively read. A PKG containing medical health information could provide a means for the patient, and their family, to give different health service providers access to pertinent information and thus facilitate appropriate diagnosis and treatment. In principle, each provider could have a pre-defined expected view or subset of the PKG that is known to be of interest to their technical specialty. For example, when coming to a new dentistry clinic the patient with such a PKG could easily share with their new dentist all the facts generated in treatment with their previous dentist.

In both scenarios, the PKG acts as a personal data storage that its owner can use to securely store data of different types, including public, private, and sensitive data, and can grant different services customized access to specific parts of this data. The PKG may be set up to access data from other sources and also to synchronize data back to these sources. Overall, the PKG is a key enabler of personalized services that are offered to its owner and to users that the owner has granted access.

[1] https://www.strava.com/.

1.2. Outline

Section 2 presents existing work from multiple fields related to the concept of PKGs. In Section 3 we introduce a terminology for characterizing PKGs, describe the central components that collectively form a PKG ecosystem, and identify the main processes that are relevant for the construction and use of a PKG. We also contrast and relate our definition of a PKG to other similar concepts. Section 4 surveys related work by categorizing them according to the terminology and aspects introduced in Section 3. In Section 5 we identify gaps between the current state of the art in the field and the requirements proposed by the scenarios described in Section 1.1, and propose directions for future work. Finally, we conclude in Section 6.

2. Related work

In this section, we discuss how personal knowledge graphs are related to knowledge graphs, personalized information access, personal information management, knowledge extraction and knowledge base population, and the Semantic Web.

2.1. Knowledge graphs

The term *knowledge graph* (KG), perhaps best known due to its popularization by Google under the name "Google Knowledge Graph" in 2012,[2] is defined in many different ways—dated both before and after 2012 (cf. Hogan et al., 2021). We choose to follow the broad definition by Hogan et al. (2021): "a knowledge graph is a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. The graph of data (a.k.a. data graph) conforms to a graph-based data model, which may be a directed edge-labeled graph, a property graph, etc."

In the literature, the terms *knowledge base* and *knowledge graph* are often used interchangeably. We consider the term *knowledge base* (KB) to be a generalization of the term knowledge graph and take this to mean a knowledge graph that does not necessarily conform to the graph-based data model.

One of the benefits of using a graph model for representing knowledge is that graph data may easily be extended and integrated with other graph datasets in a dynamic and incremental manner without the need to conform to a particular predefined schema. Additionally, it supports representing entities for which some relationships or data values are missing or not known. The versatility of the graph model also allows for representing many types of data formats and kinds of data as long as it can be encoded into nodes and edges. This allows for example schema information to be represented alongside the data, which makes it possible to manage schema information using the same tools and methods as for the graph data.

Knowledge graphs are a natural choice to use for PKGs to represent personal knowledge since such data, depending on the use case, can be highly dynamic and disparate since it is subject to frequent changes and updates; express facts of different granularity, accuracy and modality; and originate from a wide range of sources that may use different representation languages and formats.

[2] https://blog.google/products/search/introducing-knowledge-graph-things-not/.

2.2. Personalized information access

The amount of information on the Web has made it difficult and time-consuming for a user to search for specific information. The gap between how well a system could perform if it tailors results to the individual, and how well it performs by returning results designed to satisfy everyone is referred to as the *potential for personalization* (Teevan et al., 2010). It represents the potential improvement of the returned ranking to be achieved by targeting the needs of a specific individual.

Based on this idea, personalized services have been developed to tailor the information presented to a user based on the representation of their long-term interests (White, 2016). Previous studies have shown that using information such as document content or concepts, browser history, query history, and user groups enhances the ranking of documents retrieved for individual users (Matthijs and Radlinski, 2011; Sontag et al., 2012; Teevan et al., 2009; Dou et al., 2007; Shen et al., 2005). Collecting such information over a significant period of time allows for the creation of a user model that represents the user's long-term interests.

There are two main approaches to provide personalized search results based on the representation of a user's interests. On one hand, the top-n retrieved documents are re-ranked based on the user model in order to promote documents matching user's preference to the top (Sontag et al., 2012; Teevan et al., 2005), hence, increasing their chance to be inspected. On the other hand, the user model can be considered as a part of the ranking algorithm, i.e., the retrieved documents for a query are biased towards the user's preferences (Agichtein et al., 2006).

However, personalization is not an optimal solution in dynamic, biased, and data-intensive environments where the user's information needs are changing constantly. The epistemic bubbles are an example of this; these bubbles exclude relevant information, perhaps accidentally, because it does not match the user's interests (Nguyen, 2020). For example, two users who have different political positions will not likely see the same information, as personalization will filter out the information that contradicts the user's beliefs.

The question of user privacy is a recurrent issue in the field of personalized information access. Indeed, there is a tension between privacy and personalization, as the latter needs to collect user information that can reveal private information, such as political inclination and profession (Shen et al., 2007). Therefore, a compromise between how much the users agree to share and how much user information is needed to provide a personalized service should be considered (Panjwani et al., 2013). Our proposed PKG ecosystem places a strong emphasis on privacy. Indeed, the owner of the PKG has a more direct control over the data they share with each service, allowing them to decide how to balance the trade-off between privacy and personalization.

2.3. Personal information management

Personal information management (PIM) "refers to the practice and the study of the activities a person performs in order to acquire or create, store, organize, maintain, retrieve, use, and distribute information in each of its many forms [...] as needed to meet life's many goals [...] and to fulfill life's many roles and responsibilities" (Jones et al., 2017). While the origins of PIM may be traced back as far as the seminal article "As We May Think" by Bush (1945), where he describes the concept of the memex that would make knowledge more accessible, the inception of contemporary PIM dates back to the formation of a special interest group at the 2004 CHI Conference on Human Factors in Computing Systems (Bergman et al., 2004), followed by an NSF workshop in 2005 (Jones and Bruce, 2005). PIM places special emphasis on the organization and maintenance of personal information items for later use and repeated reuse (Jones et al., 2017). An *information item* is defined to be an encapsulation of information in a persistent form that can be managed (i.e., created, stored, copied, moved, deleted) (Jones et al., 2017). Examples of information items include files, emails, web pages, posts, and status updates on social media platforms. Importantly, information items are not restricted to digital forms, but can also be paper-based documents. There are several ways in which information can be *personal*: it can be (P1) controlled/owned by the individual (e.g., email messages in one's account, files on a hard drive or in a cloud service), (P2) about the individual (e.g., credit history, medical records, web search history), (P3) directed towards the individual (e.g., emails, web ads, tweet mentions), (P4) sent/posted/shared by the individual (e.g., sent emails, published articles), (P5) things experienced by the individual (e.g., web history, photos, videos), (P6) potentially relevant/useful to the individual (e.g., future job, home, partner) (Jones et al., 2017). As noted by Jones et al. (2017) "the senses in which information can be personal are not mutually exclusive". For example photos taken at a given event may be owned (P1), about (P2), shared (P4), and experienced (P5) by the same person.

PIM is associated with three main activities (Jones et al., 2017): (1) *keeping activities* include decisions concerning what subset of the encountered information and how should be kept for later use, (2) *finding/re-finding activities* include explicit searches as well as various navigation activities performed to locate information, (3) *meta-level activities* focus on connecting information with needs and involve organization, maintenance, and making sense and use of personal information. A distinctive characteristic of PIM research is its strong focus on the human perspective: identifying patterns of behavior in how people approach different forms of information using various computer-based tools. For example, recurrent themes of discussion include the use of folders versus tags (Bergman et al., 2013a; Civan et al., 2008; Voit et al., 2012) and navigation versus search (Bergman et al., 2008, 2013b; Fitchett and Cockburn, 2015; Teevan et al., 2004). Many excellent studies focus on how people use and organize specific forms of information, e.g., their email (Bellotti et al., 2003; Capra et al., 2013; Hanrahan and Pérez-Quiñones, 2015; Whittaker et al., 2011) and bookmarks (Abrams et al., 1998; Boardman and Sasse, 2004; Jones et al., 2002), but this also makes the field of PIM fragmented. PIM is also related to the notion of *quantified self*, which is concerned with the tracking of personal activities, often through a dedicated hardware device (e.g., physical fitness monitors and activity trackers such as smartwatches) (Gurrin et al., 2014).

PIM is closely related to the notion of PKGs, but there are several key differences:

- PIM has a strong emphasis on human activities around managing personal information, i.e., the human–computer interaction is in focus. PKGs center around the information itself, how it can be represented and utilized across services and applications.
- The atomic units in PIM are information items, while PKGs operate on facts, which is a finer granularity. Also, PKGs assume digital information, while PIM in the broader sense also includes paper-based documents.
- Underlying all PIM activity types is an *implicit* effort to make sense of the available information. In PKGs, sense-making is *explicit* in the representation of information as facts.
- All PIM activities are driven by the aim to assist in satisfying the user's *information needs*. PKGs have a broader scope. For instance, in Example 1, the personal assistant can proactively take initiative, without addressing any existing information need. Another use of PKGs is to share data with others, as illustrated with the case of health service providers in Example 2.
- Integration is part of both PIM and PKGs, but for different reasons: in PKGs it allows for providing better services, while in case PIM it helps to counter information fragmentation.

PKGs can be leveraged in PIM for organizing personal information in a finer granularity (i.e., as facts as opposed to files/documents). In our envisaged PKG ecosystem, privacy and access management would be taken care of, and PIM could focus on building personalized services that utilize or help maintain this information.

2.4. Knowledge extraction and knowledge base population

A wealth of information resides in unstructured or semi-structured format (text documents, social media posts, multimedia files, etc.) that is not readily available in knowledge bases. Knowledge bases can be augmented by extracting structured information from these sources. *Knowledge acquisition* (a.k.a. *knowledge harvesting*) refers to the process of extracting information on entities and relationships from a large data corpus (e.g., the Web) and turning them into machine-readable facts (Weikum et al., 2021).

Knowledge base population is a more specific task within the broader problem space of knowledge acquisition, focusing on the augmentation of an existing KB with entities, types, attributes of entities, and relationships between entities (Ji and Grishman, 2011). A key component of KB population is *entity linking*, which is concerned with the resolution of entity mentions detected in text to unique identifiers in a KB; it is typically performed by leveraging contextual information and existing knowledge bases (Balog, 2018). It is important for a KB to have a clean and expressive taxonomy of types (classes) and that these are populated with uniquely identified entities. *Class-instance acquisition* aims at obtaining additional entities that belong to a given entity type (a.k.a. class) (Pantel et al., 2009) and obtaining additional entity types for a given entity (Gangemi et al., 2012). The process of populating the KB with new facts about entities, i.e., additional attributes (with literal values) and relationships (with other entities) is often referred to as *slot-filling* (Ji and Grishman, 2011). Extraction is traditionally approached using pattern- and rule-based techniques. Restricted kinds of patterns may be learned automatically from examples (Sarawagi, 2008). More recently, the task is viewed either as a classification or as a sequence-tagging problem (Weikum et al., 2021).

What we have discussed above, populating an existing KB with additional facts about entities, is an instance of *targeted* and *closed* information extraction. Targeted, because it focuses on the extraction of a predefined set of predicates (attributes and relationships) for specific entities, and closed in a sense that all entities, types, and relationships already have canonicalized (unique) identifiers (Balog, 2018). Open information extraction addresses the discovery of new attributes and relationships and their organization into a canonicalized format with clean type signatures (Weikum et al., 2021). *Novel entity detection* deals with the discovery of out-of-KB entities; it involves identifying and classifying previously unseen or unknown entities within a given text (Lin et al., 2012). *Class-attribute acquisition* refers to the discovery of relevant attributes (and types of relationship) of classes (Pasca and Durme, 2007). *Predicate discovery* is the task of extracting predicate-argument structures, where the predicate and two or more arguments take the form of short natural language phrases extracted from an input sentence (Weikum et al., 2021). Traditionally, most open information extraction methods relied on patterns and rules (Fader et al., 2011). More recently, neural approaches have been proposed, following the success of deep learning models on various NLP tasks (Cui et al., 2018). However, canonicalization of the resulting facts for inclusion in a KB, especially of the newly seen phrases describing predicates, remains an open challenge (Weikum et al., 2021).

2.5. The Semantic Web and Semantic Web technologies

The Semantic Web (Berners-Lee et al., 2001) is an extension of the World Wide Web (the Web) that combines standards for knowledge representation with established web standards and architecture with the aim of making data available on the Web machine readable and "understandable". This is done by using formal languages with defined syntax and semantics, such as the Resource Description Framework (RDF)[3] and the Web Ontology Language (OWL),[4] to encode the data and its semantics. The original vision of the Semantic Web is expressed by Tim Berners-Lee as one where computers "become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A 'Semantic Web', which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines". In the seminal paper "The Semantic Web" (Berners-Lee et al., 2001), this is illustrated by an example reminiscent of our motivating scenarios: in the example, Pete and Lucy's (personal) *Semantic Web agents* together organize and schedule a series of health treatments for their mom so that they fit her insurance plan, minimize traveling time, and fit with their own schedules. The agents are able to read and understand the prescribed treatment from their mother's doctor, listings of specialist providers, map data and personal calendar data, and to communicate with each other autonomously and with their users. It is clear that the vision of The Semantic Web greatly overlaps with the expected outcomes of using PKGs. However, their goals differ in that that personal aspect of PKG is its very definition, while The Semantic Web has the much broader scope of enabling more sophisticated machine to machine communication based on the explicit representation of the semantics of data.

The Semantic Web community has developed a suite of technologies that are relevant for PKGs. The Resource Description Framework (RDF)[5] is one of the most important technologies in the *semantic web stack*[6] and is a prominent example of a directed edge-labeled graph model (Hogan et al., 2021). An *RDF graph* is a set of *RDF triples* $\langle s, p, o\rangle$ where the *predicate* p is the labeled edge element that connects the *subject* element s to the *object* element o. Sometimes in the literature, the term *SPO-triples* is used to designate RDF triples. The elements of a triple are, with some restrictions, either an *IRI* (Internationalized Resource Identifier), a *literal*, or a *blank node*. Well-known KGs that use RDF are DBpedia (Auer et al., 2007) and Wikidata (Vrandečić and Krötzsch, 2014). RDF KGs are often published following principles known as Linked (Open) Data (Heath and Bizer, 2011) that promotes sharing and connecting data to support machine-readability over the Web using established Web standards such as HTTP (Hypertext Transfer Protocol), IRIs, and RDF. Additionally, RDF KGs are often made available via SPARQL endpoints through which the data may be accessed and queried using the query language and protocol SPARQL.[7] The Web Ontology Language (OWL) is the de facto standard for representing logical ontologies and is used to define the vocabulary and semantics of the schema used by the data.

3. Personal knowledge graphs

This section presents our definition of a PKG and discusses how it differs from existing ones (Section 3.1). Subsequently, the ecosystem within which the PKG exists is introduced and its main aspects are presented (Section 3.2).

3.1. What is a personal knowledge graph?

In prior work, the concept of a PKG is often only implied rather than explicitly stated, leaving room for different interpretations. Below, we present the definitions for the various interpretations that exist, starting with our proposed definition. The key differences between the various interpretations are summarized in Table 1.

[3] http://www.w3.org/TR/rdf11-concepts/.
[4] http://www.w3.org/TR/owl-overview.
[5] http://www.w3.org/TR/rdf11-concepts/.
[6] http://www.w3.org/2000/Talks/1206-xml2k-tbl/.
[7] http://www.w3.org/TR/sparql11-overview/.

Table 1
Different interpretations of personal knowledge graphs.

	PKG (this paper) (cf. Definition 1)	PKG (Balog and Kenter, 2019) (cf. Definition 2)	Personalized knowledge graph (cf. Definition 3)
Ownership	Created and maintained by an individual	Created and maintained by an individual	Created and maintained by a service
Public facts	Can incorporate facts from public knowledge graph	Public facts are not explicitly stored, but can be linked	Built with facts from a public/proprietary knowledge graph
Private facts	The owner of the PKG can add private facts (e.g., beliefs) as long as they have the correct format	The owner can add private facts (e.g., beliefs) as long as they are connected to it	Facts that are not of public knowledge cannot be stored (e.g., an individual medication regimen)
Graph structure	Facts do not need to be connected to the user	All facts in the PKG are connected to the user resulting in a spiderweb layout	Facts do not need to be connected to the user

Definition 1 (*Personal Knowledge Graph*)**.** A *personal knowledge graph* (PKG) is a knowledge graph (KG) where a single individual, called the owner of the PKG, has (1) full read and write access to the KG, and (2) the exclusive right to grant others read and write access to any specified part of the KG. The primary purpose of the PKG is to support the delivery of services that are customized particularly to its owner.

Note that we do not pose any requirements to the contents of the PKG; the discriminating factor from regular KGs is the administrative rights to the KG that acts as a PKG. This is different from the definition by Balog and Kenter (2019), which has enjoyed a wide adoption within the research community (Tiwari et al., 2023):

Definition 2 (*Personal Knowledge Graph Balog and Kenter, 2019*)**.** A personal knowledge graph is a source of structured knowledge about entities and the relations between them, where the entities and the relations between them are of *personal*, rather than general, importance. The graph has a particular "spiderweb" layout, where every node in the graph is connected to one central node: the user.

Rather than requiring that the owner is explicitly represented in the PKG and that all facts in the PKG are connected to the owner, our definition establishes this relation though the administrative rights to the PKG; as the facts in the PKG can only exist in the PKG by the owner's discretion, they are by definition also personal.

In the literature, we observe that a majority of the applications focus on the personalization of a service. For example, Lu et al. (2019) and Zhang et al. (2018) propose solutions for a dialogue agent to provide personalized answers to a user. To address this problem, it is common to create a representation of a user such as a user profile. This representation can take the form of a KG which is sometimes referred as a PKG. However, we argue that they actually refer to a *personalized knowledge graph* (a.k.a. *personal interest graph*). Rastogi and Zaki (2020) distinguish between personalized KGs and PKGs at the level of stored information, with personalized KGs limited to the entities described in the general KGs and PKGs complementing general KGs with additional, personal information about the user.

Definition 3 (*Personalized Knowledge Graph*)**.** A *personalized knowledge graph* is a subset of an existing knowledge graph, restricted to entities and relationships that can characterize the interests of a given individual.

In the case of a personalized knowledge graph, the user rarely knows how it is created and with which facts it is populated. Thus, it does not fulfill the ownership criterion of our definition.

3.2. The PKG ecosystem

As one of the main contributions of this work, we emphasize the necessity of situating the PKG within its broader operational ecosystem in order to conduct meaningful research. Our proposed unifying architecture for PKG ecosystems is presented in Fig. 1. We identify three main aspects: *population*, *representation and management*, and *utilization*, all of which are required for the successful engineering and operation of a PKG.

3.2.1. Population

Population concerns the aspect of adding data to the PKG from existing data sources or services (cf. Sections 2.3 and 2.4). We consider three different types of data sources:

- *Private data sources* refer to data in any format that are private to the PKG owner, meaning that the owner typically is the only one with the access to read and write to the contents of these sources.
- *Public data sources* refer to data in any format that are publicly available.
- A specific type of public data sources in the PKG ecosystem are public KGs, often called *Linked Open Data sources* (Heath and Bizer, 2011). These data sources are assumed to be in a format that is already compatible with the format of the PKG. That is, they may be integrated with the PKG without any additional preprocessing or alignment steps, which would typically be required when populating data from the two other types of data sources. Rather, the PKG and Linked Open Data sources both link to vocabulary definitions in external ontologies.

The population aspect also includes the process of synchronizing any modifications made to the extracted data in the PKG back to its original data source. This is only relevant for private data sources as these are the only type of sources we assume the owner has access to update.

Example 1 (*continued*)**.** In the Personal Trainer Assistant example, the PKG is populated with private calendar data from the PKG owner's calendar, data about the owner's sport interests from Facebook and YouTube, and public health related Linked Open Data from Wikidata. The relevant data from these sources must be extracted, represented and formatted to fit with the representation requirements of the PKG.

The population aspect of the PKG ecosystem interfaces with the representation and management aspect, and it is important that the data output from the population processes respect the requirements set by the representation aspect.

3.2.2. Representation and management

This aspect is concerned with *representation*, i.e., the logical representation, format and expressivity of the facts and statements that make out the contents of the PKG, and *management*, i.e., the organization, storage, retrieval, and access control of the contents of the PKG, whose functionality is collectively made available by the *management system* of the PKG.

The PKG contents should be organized to enable efficient and accurate data capture, access, and update. A PKG should be able to store and organize a wide range of facts or statements, e.g., objective facts ("the capital of Norway is Oslo"), records of personal events ("I visited the dentist May 1st 2020"), logging ("I accessed the medical records in my PKG on May 2nd, 2020"), but arguably also more complex statements such as personal beliefs and probabilities ("I believe the Earth is flat") and other person's beliefs ("My mom thinks my dentist is very polite"). To support this range of types of statements, a rich set of metadata and context data becomes vital: where, when, and how did the statement

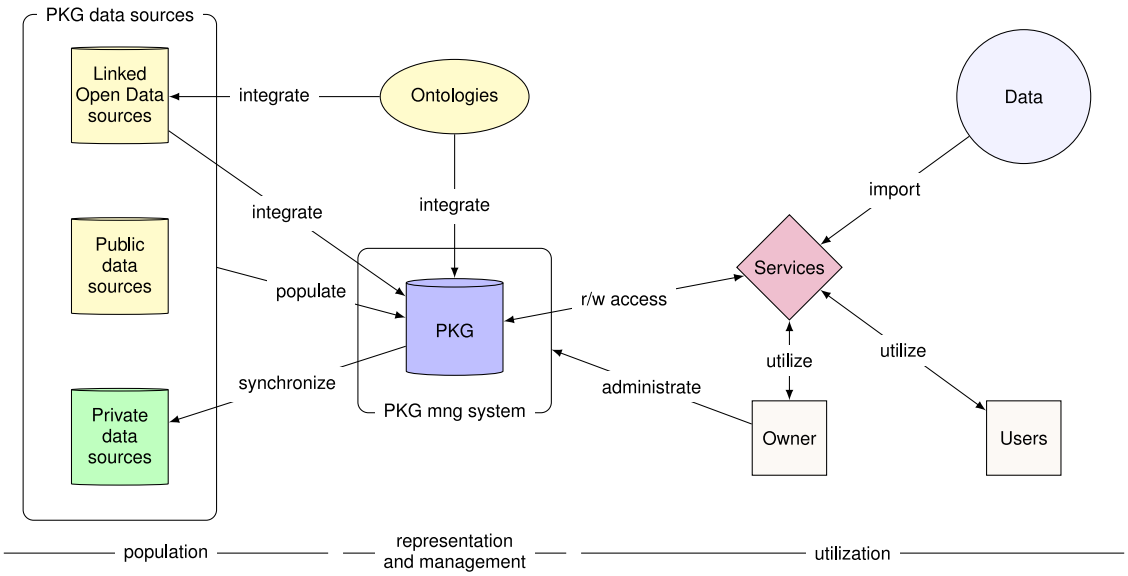

Fig. 1. PKG ecosystem.

come about, and who holds the fact to be true and to which degree. This will allow the services that access the PKG data to better assess, select, and use relevant data. Provenance data is also vital for enabling synchronization with private data sources. It is reasonable to think that the size of the metadata will greatly exceed the size of the "actual data".

Furthermore, the data in the PKG should be semantically rich so that services that access the PKG can exploit the semantic descriptions, metadata, and context descriptions in order to combine PKG data in new ways and to saturate the PKG data with new knowledge that will improve the data foundation for the services. This requires the PKG to integrate with several ontologies that can capture the semantics of the facts stored in the PKG.

Example 1 (*continued*)**.** In order to efficiently and correctly manage the data collected for the Personal Trainer Assistant service, the PKG management system will benefit from a detailed and accurate set of provenance data. For instance, when suggesting training exercises, a preference explicitly set by the owner should carry more weight than a preference inferred from the owner's YouTube viewing history. Additionally, recent data records should take priority over older records. By connecting the owner's training preferences to rich semantic descriptions in ontologies and KGs, e.g., ontologies about the human anatomy or KGs about exercises and their benefits, the collected PKG data can be saturated with more data inferred from reasoning over the combination of the collected data and the integrated ontologies. For example, a service could infer by reasoning over ontologies of the human anatomy and exercise injuries that the current runner's knee problem suggests avoiding sport activities similar to running.

Example 2 (*continued*)**.** In our Sharing Health Information scenario, assuming that the PKG contains the complete medical history of the owner, a PKG service could monitor the drug prescriptions given to the owner, and by reasoning over drug and medicine ontologies discover and flag new prescriptions that are incompatible with previously given prescriptions or with any of the owner's recorded illnesses and injuries.

The PKG management system is responsible for handling the storage and retrieval of data contained in the PKG, managing and controlling access to and the security of the PKG, and providing the administrative management service to the owner of the PKG. Using the management system, the owner should be able to maintain the contents of the PKG, and grant read and write access to the PKG for other services and users.

Example 2 (*continued*)**.** The PKG management system plays an important role in organizing access for different users to different parts of the PKG data: the owner should have read and write access to the complete contents of the PKG, while the owner's dentist should only have read access for the data content that is relevant for the owner's treatment.

Note that by definition the owner has read and write access to the PKG; manipulating certain type of information, such as medical records, can have serious negative consequences. We come back to this particular issue later, in Section 5.

3.2.3. Utilization

Utilization is the task of exploiting the PKG data to deliver successful personalized services to its owner and users, and this is where the real value of the PKG is realized. The services can be personalized only because they access the PKG, but they may also interface with other data sources—in the widest sense, e.g., weather data or Twitter feeds—to deliver value to their users. These services must be able to correctly interface with the PKG management system and understand and exploit the potentially very rich structure of the PKG data.

Example 1 (*continued*)**.** The Personal Trainer Assistant service could take the form of a mobile phone application. The service would connect to and communicate with the owner's PKG to access the relevant data. The PKG management system would control and manage this access, and handle any queries or other requests issued to the PKG by the Personal Trainer Assistant service, such as requests for reasoning over the data with respect to specific ontologies or requests for access to more of the PKG's data. The service can also communicate with other data sources than the PKG, e.g., news channels or Twitter feeds, including data that the vendor of the service application makes available for this particular use, such as news about training trends or sport results, or personalized suggestions made by the service vendor's personnel. The service could then use this data and match it with the relevant data in the PKG to make customized suggestions to the owner. The results would then need to be presented to the owner in the application in a user friendly manner.

4. Survey

In this section, we present previous work on personal knowledge graphs by mapping it to our PKG ecosystem (shown in Fig. 1). For

Table 2
Paper categorization based on PKG ecosystem; ● denotes well described process , ○ denotes briefly described process.

	Population		Representation and management			Utilization	
	Populate	Synchronize	Represent	Integrate	Manage	Administrate	Utilize
Groza et al. (2007)	●	○	●	●	●	●	●
Sambra et al. (2016)	○	●	●		●	●	●
Gyrard et al. (2018)	●		○	○			○
Mazaré et al. (2018)			○				○
Zhang et al. (2018)			○				○
Lu et al. (2019)			○				○
Luo et al. (2019)	●		●				●
Yen et al. (2019)	●		●	●			○
Tigunova et al. (2019)	●		○				○
Gerritse et al. (2020)							○
Rastogi and Zaki (2020)	○		○	○			○
Tigunova et al. (2020)	●		○				○
Ammar et al. (2021)	●		●	○	●	●	●
Vannur et al. (2021)	●		○	●			○
Seneviratne et al. (2021)	●		●				●
Chakraborty et al. (2022)	●		●		●	○	○

each work, we identify the aspects and processes that are studied; Table 2 presents a summary of our findings, evaluating how well each surveyed work addressed each of the processes identified in our PKG ecosystem framework.[8] Below, we organize the discussion according to the main aspects, i.e., population, representation and management, and utilization, in respective Sections 4.1–4.3. Within each section, studies are presented in chronological order. Where applicable, we describe the surveyed works with the same terms as in our PKG ecosystem, while also respecting each work's terminology where something more specific or nuanced is being expressed.

4.1. Population

We first consider previous work in terms of how external data is used to populate a PKG, and how any update in the PKG is synchronized back to the external data sources. The facts used to populate a PKG can come from both public (e.g., food calorie charts, public healthcare policies) and private (e.g., emails, medical records) data sources.

Groza et al. (2007) present the NEPOMUK project, an architecture of a social semantic desktop, which is a tool to share data between a local user and other users or applications, inspired by semantic web technologies. In NEPOMUK, the PKG, which is only implicitly defined, can be manually populated with RDF data. In addition to manually inserting data, services such as Data Wrapper and Text Analysis are intended to help populate the PKG with application data (email, calendar) and unstructured data (e.g., free text in documents). The synchronization is enabled only for the resources (e.g., documents) shared with collaborators within one closed group and it does not include propagating PKG changes to the private data sources. However, the described scenario with notifications can be a first step towards the implementation of the synchronization process.

The Solid (Socially Linked Data) project introduced by Sambra et al. (2016) is a protocol based on Semantic Web technologies with the purpose of re-decentralizing the Web. Solid allows users to own their personal online data stores (PODs), which is similar to our notion of a PKG, and control applications' access to operate on these PODs. Solid does not offer an automatic population of a POD, in particular not from unstructured sources. Synchronization is enabled via the implemented PubSub system based on WebSocket,[9] which allows the POD to send live updates and notifications regarding resources of interest.

Gyrard et al. (2018) present a personal health knowledge graph (PHKG) that represents medical and personal data to support the development of a personal health coach or digital health advisor. The PHKG is populated with data from the kHealth project datasets and Linked Open Data resources. Additionally, the Kno.e.sis Alchemy API[10] is used to extract structured data from different sources such as clinical trials.

Luo et al. (2019) investigate goal-oriented dialogue systems which harness models of both (1) the personality and language style of the user, and (2) the preferences of a user with respect to the system's underlying knowledge graph. The proposed dialogue system profiles the personality of the user by adding personal information terms, such as gender, age, and diet, for each dialogue turn in a slot-filling manner.

Yen et al. (2019) study the extraction of life events from life logs such as tweets to populate a personal KB for memory recall and life support assistance applications. The population of personal KB is not limited to life events and includes entities from DBpedia (Auer et al., 2007) and Freebase (Bollacker et al., 2008).

The methods for inferring user attributes from a sequence of utterances presented by Tigunova et al. (2019) are used to create a personal KB that can serve as a distant source of knowledge for personalization. The authors propose to use deep learning (Hidden Attribute Models) for inferring personal attributes, such as profession, age, or family status, from conversations (Reddit discussions, movie scripts, and crowdsourced personal dialogues).

The review presented by Rastogi and Zaki (2020) covers personal health knowledge graphs (PHKG) for patients. PHKG is defined as a representation of aggregated multi-modal data including all the health-related personal data of a patient. It can be generated by inferring patient preferences over a given general-purpose KG (the authors mention the usage of entity linking).

The approach presented by Tigunova et al. (2020) (built on top of work presented in (Tigunova et al., 2019)) proposes CHARM (Conversational Hidden Attribute Retrieval Model), a zero-shot learning method that creatively leverages keyword extraction and document retrieval in order to predict attribute values that were never seen during training. CHARM can be used for inferring attribute values in a zero-shot setting and extracting personal information from conversational utterances to populate the personal KB. The details of populating the personal KB with extracted attributes are not provided.

[8] To illustrate how previous works are summarized in Table 2, consider the "populate" column. If a work does not address the "populate" process, then that cell of the table is left blank. If the "populate" process is acknowledged as a part of the PKG ecosystem, e.g., the importance of extracting facts from data, but no implementation details are mentioned, then that cell of the table is marked with ○. Finally, if the work addresses the process in practical terms, e.g., as many of the surveyed works research the actual process of extracting facts from data, then that cell in the table is marked with ●.

[9] http://www.w3.org/TR/websockets/.

[10] http://wiki.knoesis.org/index.php/Knoesis_Alchemy_of_Healthcare.

Ammar et al. (2021) study requirements and current shortcomings of electronic health records and use the Solid (Sambra et al., 2016) platform to propose and prototype personal health libraries, and a mobile app using data from them. The personal health libraries are to be populated using existing Semantic Web technologies, such as Linked Open Data and the Web Annotation Data Model.[11] The population process follows the REST principles. In addition, text summarization and knowledge mapping are mentioned as future directions to be explored in populating the PKG.

Vannur et al. (2021) discuss the subtasks of knowledge base population applied to extracting entities and relations from free text, specifically for the purpose of personal knowledge base population. It is defined as populating a knowledge base with personal information, albeit not necessarily limited to a PKG in the sense used in the present work. The focus of the paper is the subtask of extracting facts from free text.

The personal health ontology used to generate PHKG (Gyrard et al., 2018) is presented by Seneviratne et al. (2021) on the use case of diet recommendation. The authors use time series summarization to extract RDF triples from food logs that are then used to populate the PHKG. The theoretical considerations on modifications required in an existing framework to perform this, as well as a specific example are provided. However, implementational details are missing.

Chakraborty et al. (2022) describe the use of PKGs to support academic researchers in their individual and collaborative research activities. The personal research knowledge graphs presented in the paper can be populated with the entities and relationships between them that are automatically extracted from different sources such as scholarly papers. Both manual management and automatic extraction from unstructured data using NLP techniques are mentioned. As the authors follow the definition of Balog and Kenter (2019), entities present in public knowledge bases (e.g., Wikidata and Open Research Knowledge Graph[12] Jaradeh et al., 2019) are not integrated but linked.

Summary. The majority of the papers discussed above provide broad and detailed descriptions of the PKG population aspect. Several works focus only on the problem of extracting information from existing data sources, e.g., (Tigunova et al., 2019, 2020) and adding them to the PKG, e.g., (Gyrard et al., 2018). In these works, extraction is typically restricted to a selected set of predicates (Tigunova et al., 2019, 2020; Yen et al., 2019), or is not stated explicitly. Interestingly, only two works mention the synchronization aspect (Groza et al., 2007; Sambra et al., 2016). These papers, inspired by Semantic Web technologies, propose standalone architectures for personal online datastores (Sambra et al., 2016) or sharing data between users and applications (Groza et al., 2007).

4.2. Representation and management

The way information is represented in a PKG, how data from external data sources are integrated with the ontology of the PKG, as well as the details of the concrete software used to manage the PKG, together comprise the second aspect of the PKG ecosystem. Here, the management system is considered without the elements of administration interfaces which the management system nevertheless enables, which are discussed in Section 4.3.

In the NEPOMUK project, Groza et al. (2007) define a data resource as an RDF graph obeying some ontology or set of ontologies. To integrate external data, a Mapping Service translates RDF graphs from a source ontology to a target ontology. Groza et al. (2007) also discuss access rights management and access control implemented with the NEPOMUK middleware, and mention the use of the Web Services Description Language (WSDL) to define services.[13] The work further envisions a shared information space based on peer-to-peer exchange, and consider both local and distributed storage of data.

Solid (Sambra et al., 2016) represents structured data as RDF, while unstructured data may be of any type, e.g., images, video, or free text. Application data is stored in documents with individual IRIs for each resource. Solid specifies requirements for the personal online datastore (POD) management system and Sambra et al. (2016) offer several Solid prototype servers to explore this specification. The management system must address RDF and non-RDF resource storage, basic data operations (Linked Data Platform[14] operations and some Solid extensions of these), access control, and, optionally, complex data retrieval. Specifically, the users' data are stored in PODs and managed in a RESTful way, with SPARQL support for complex data retrieval operations.

Gyrard et al. (2018) describe the personal health knowledge graph (PHKG) in terms of component technologies, including the use of existing medical ontologies, but do not explicitly address the representation or management system of the PKG for health. They do discuss the integration of public ontologies and KGs from ontology catalogs such as BioPortal[15] and Linked Open Vocabulary.[16] They also mention some challenges in reusing existing ontologies, but do not share how these challenges are tackled in practice.

Mazaré et al. (2018) present a large dataset of persona-based dialogues built using conversations extracted from Reddit, as well as end-to-end dialogue models trained on this dataset. The conversational dataset is used to extract a persona for some users, where a persona is represented as a set of sentences representing the personality of the responding user.

Zhang et al. (2018) present a dataset of chit-chats with personas and models trained on this dataset, conditioning next utterances on personas (for either or both sides of the dialogue) to be more engaging. Here the persona is represented as a set of maximum five sentences describing the persona in the first person.

Lu et al. (2019) investigate different conversational agents to provide personalized customer service chat by exploiting customer profile information. The profile used for personalization is represented only by a few facts such as "customer's membership status, the order fulfillment method, the shipping carrier, whether the order is a single or multi-item order, and whether the order was eligible for cancellation at the time of contact". The profile information is not controlled by the person who is being profiled.

Luo et al. (2019) first express the personal profile of the user as a concatenation of one-hot vectors, each representing a selected user attribute and its value. The user's profile model and preference model are then represented as neural embeddings to support the ranking of knowledge base items.

Yen et al. (2019) extract life events that are represented with quadruples of the form (object, predicate, subject, time). The quadruples are inserted into a knowledge base. The extraction of life events is integrated with predicates from the Chinese FrameNet ontology (Yang et al., 2018).

Tigunova et al. (2019) mention a personal KB without revealing any details on how it is constructed using the extracted attributes. However, the attributes are represented in the form of SPO triples.

Rastogi and Zaki (2020) represent data using classes, entities, attributes, and relationships. They mention integration with public KGs and the use of predefined ontologies without giving any implementation details.

Tigunova et al. (2020) represent extracted attributes in the form of attribute-value pairs. Here, the personal attributes of interest are only

[11] https://www.w3.org/TR/annotation-model/.
[12] https://www.orkg.org/orkg/.
[13] https://www.w3.org/TR/wsdl/.
[14] https://www.w3.org/TR/ldp/.
[15] https://bioportal.bioontology.org/.
[16] https://lov.linkeddata.es/dataset/lov/.

"profession" and "hobby", and the possible values are drawn from the corresponding Wikipedia "List of" pages.[17]

Ammar et al. (2021) directly adopt the Solid approach and apply it to represent personal health data as various RDF and non-RDF data, distributed over potentially multiple PODs per patient/user, with a primary focus on the patient/user's RDF-based PKG of health information. They also mention knowledge mapping and the use of public knowledge bases and ontologies, but leave integration implementation details for future work. The management system is likewise directly derived from the Solid platform.

Vannur et al. (2021) are motivated by the problem of populating a PKG, but focus on extracting entities and relations. The exact representation of extracted information for PKG population is only discussed speculatively as the graph data could be exported to different formats, including RDF. The authors consider public KGs such as YAGO[18] and the Person Ontology (Ganesan et al., 2020). However, in practice they address integration by training models for entity classification and link prediction on existing datasets (OntoNotes[19] Hovy et al., 2006, TACRED[20]) with their respective attributes and entity types.

Seneviratne et al. (2021) propose a new ontology for their PHKG, the Personal Health Ontology. We note that their ontology reuses concepts from existing ontologies such as the Statistics Ontology.[21] To the best of our knowledge the Personal Health Ontology is not publicly available. RDF triples are used to represent the data.

Chakraborty et al. (2022) suggest that a Personal Research Knowledge Graph (PRKG) may be modeled as a labeled property graph in Neo4j,[22] which may be serialized as an RDF graph for applications. Neo4j may be used without a predefined schema, and this is beneficial in the scenario of ongoing discovery. The facts in the PRKG are represented with SPO triples. The use of public knowledge graphs is mentioned, but integration such as ontology mapping is not explored. The management system is described only lightly, but some details may be implied by the choice of a Neo4j property graph implementation. Chakraborty et al. (2022) do address access control, and find that role-based access control (Ferraiolo et al., 2003) on a node- or relation-level is not currently implemented in Neo4j, which they intend to correct.

Summary. The papers detailed above exemplify different approaches to the aspect of representation and management in the PKG ecosystem. The surveyed research is primarily concerned with extracting facts from unstructured data such as free text or chat dialogues. These are then to be inserted into a KG, where the final representation of each fact is expected to be SPO triples, and these facts are sometimes informed by pre-existing ontologies. This may be limited to conforming extracted predicates and entities to a pre-existing vocabulary. An interesting variation is given by Yen et al. (2019), who consider quadruples that extend the SPO format with an element of time. However, past research often describes the extraction of facts from unstructured data without explicitly addressing the format of the KG. In addition, some research on personalized KGs considers attribute-value pairs, which may imply that the person being profiled is the subject with extracted attributes and values as predicates and objects, respectively. Finally, Mazaré et al. (2018) and Zhang et al. (2018) describe the use of full natural language sentences to personalize dialogue agents. Broadly, the research surveyed here is primarily concerned with other aspects of the PKG ecosystem than representation and management, with implementation details of representation of knowledge, integration with ontologies, and the management software of the PKG largely relegated to either future work or defaulted to RDF format. Groza et al. (2007), Sambra et al. (2016), and Ammar et al. (2021) comprise exceptions to this rule, and present more complete systems in terms of our PKG ecosystem.

4.3. Utilization

Finally, we survey work in terms of the utilization of the PKG by its owner or by external services. More specifically, in our ecosystem, the owner of a PKG can administrate it (e.g., add new facts manually, grant access to external services or users), and external services can interact with the PKG depending on their access privileges.

In the platform proposed by Groza et al. (2007), the NEPOMUK API is intended to connect services to a semantic desktop that facilitates data sharing between the local user and other users and services. In particular, publish/subscribe services defined by SPARQL queries enable information streams to share updates with a community of users or applications. Furthermore, in the NEPOMUK platform, resources and the corresponding RDF descriptions can be added manually. The user also decides which information sources and users can be trusted. The access control system limits the individual user's actions in the shared information space of the community. For example, a user can share a personal document only to a specific group of users.

Solid (Sambra et al., 2016) specifies access control using WebID[23] to identify individuals and groups, and the WebAccessControl[24] ontology to describe access permissions to different resources for different WebIDs. The process to administrate the personal online data stores (PODs) is not described in detail, but would have to be implemented using the Linked Data Platform and Solid operations such as those used in updating or querying PODs. In Solid, the user may control applications' access to personal data, and a consistent API for the PODs supports interoperability and the users' freedom to switch between similar applications. Sambra et al. (2016) present a number of typical applications, such as a contact manager, to show that Solid offers integration with multiple social Web applications for common day-to-day tasks.

Gyrard et al. (2018) argue that a personal health knowledge graph (PHKG) is well suited for a personalized health coach application. More particularly, they discuss the use-case of self-management of chronic diseases such as asthma. Later work by Rastogi and Zaki (2020), Ammar et al. (2021), and Seneviratne et al. (2021) also study use-cases related to the idea of a personalized health coach. Rastogi and Zaki (2020) argue that PHKGs can be used to personalize recommendations from food platforms to encourage a healthy life style. Ammar et al. (2021) exemplify the purpose of their proposed personal health libraries with a mobile app to support users' chronic disease self-management, and plan to expose the personal health libraries to third-party applications. Unlike the other work related to PHKGs in this survey, Ammar et al. (2021) discuss the administrate process by describing access control and the decoupling of data from applications accordingly. Seneviratne et al. (2021) present the use-case of an application giving personal insight for Type 2 Diabetes self-management following clinical guidelines. They propose to make recommendations by reasoning over the PHKG containing clinical guidelines represented using OWL.

Several works (Zhang et al., 2018; Lu et al., 2019; Luo et al., 2019; Mazaré et al., 2018; Tigunova et al., 2019, 2020) motivate the use of PKGs to create personalized dialogue systems, also referred as conversational agents. For example, Lu et al. (2019) propose to use profile information to adapt the behavior of a customer service conversational agent. In their work, Luo et al. (2019) describe a dialogue system using a personalized profile where neither the extracted attributes and neural embeddings representing the user, nor the knowledge graph of items for

[17] https://en.wikipedia.org/wiki/List_of_professions and https://en.wikipedia.org/wiki/List_of_hobbies.
[18] https://yago-knowledge.org/.
[19] https://catalog.ldc.upenn.edu/LDC2013T19.
[20] https://nlp.stanford.edu/projects/tacred/.
[21] https://stato-ontology.org/.
[22] https://neo4j.com/.
[23] https://www.w3.org/2005/Incubator/webid/spec/identity/.
[24] https://github.com/solid/web-access-control-spec.

the system to recommend are under the user's control. Tigunova et al. (2019, 2020) state that a personal knowledge base can be leveraged for personalization in downstream applications such as web-based chatbots and agents in online forums, however, additional details on "how" are not provided.

Yen et al. (2019) mention that their personal knowledge base with life events can be used for memory recall and living support assistance. Specifically, they discuss the use of the personal knowledge base for question answering.

Gerritse et al. (2020) study potential biases introduced by the use of PKG in conversational search. The authors discuss how and why the PKG can amplify biases in personalized services such as confirmation bias, e.g., when a user searches for a conspiracy theory, and data bias related to the population of the PKG.

Vannur et al. (2021) assume a different scenario than our user-controlled PKGs scenario, but do motivate their work with the utilization of knowledge graphs containing personal information by enterprise application services to support data protection, fraud prevention, and business intelligence.

Chakraborty et al. (2022) propose using personal research knowledge graphs (PRKG) for personalization in different scholarly applications like academic search engines and recommendation systems by sharing with them the user's personal data from the PRKG. Along the paper, the authors use the example of a researcher and how a personal assistant could benefit from the PRKG, as for the exploration a specific research space. With regards to administration, the owner has the possibility to add facts to the PRKG manually and can also check the ones added automatically. The user can also control how its personal data is shared. However, the authors do not discuss the interface between the owner and the PRKG.

Summary. Based on the papers discussed in this section, we observe that the administrative service allowing the user to directly interact with its PKG for access control or data update is largely disregarded. Indeed, only three papers provide details about it, two of which are from the field of Semantic Web technologies (Groza et al., 2007; Sambra et al., 2016). In terms of use, most of the selected papers focus on the creation of services exploiting the PKG to propose a personalized experience for each users. The application domain varies from health to scholarly work through customer service. Similarly, the type of services is diverse including conversational assistance and enterprise services. This illustrates the generic aspect of PKGs and their potential for many future applications, however, these should be aware of potential drawbacks such as biases (Gerritse et al., 2020).

5. Challenges and opportunities

The previous section has synthesized existing work on various aspects of PKGs. The main high-level observation that can be drawn from this survey is that few works focus on multiple aspects and none consider the PKG ecosystem as a whole. Therefore, we start this section with a consideration of challenges and opportunities around the PKG ecosystem in Section 5.1, followed by the discussion of open issues around the individual aspects of population, representation and management, and utilization in Sections 5.2–5.4, respectively. In order to illustrate our ideas in a more tangible and concrete manner, we will utilize the running examples introduced at the outset of the paper. Specifically, Example 1 with the personal trainer assistant will be used to elucidate near-term opportunities, while Example 2 around the sharing of health information will aid in the illustration of longer-term challenges.

5.1. Ecosystem

It is clear from the previous section that PKGs have been an active area of research, with new ground broken in various aspects. Some of these aspects are in isolation already well-established research areas, and some have been developed specifically for PKGs. However, these tasks remain underdeveloped from a holistic perspective in the context of the PKG ecosystem. Thus, a main overarching challenge remains: *How do all the components within the PKG ecosystem fit together and how should they interact with each other?*

One of the main contributions of this work has been the treatment of the various aspects, components, and processes involved in the holistic realization of PKGs, and their organization in a unified architecture (cf. Fig. 1). We have also cast prior work within our framework, which illustrates its applicability. At the same time, it is important to emphasize that it is an abstract, conceptual architecture. Let us consider next what the creation of the Personal Trainer Application would entail in practice.

Example 1 (*continued*)**.** A straightforward solution would be to have a cloud-based back-end that stores the PKG and performs the various data operations (population, synchronization, etc.). The mobile and web front-ends would fetch data from the back-end, display recommendations, send notifications, and allow the user to configure the service (features and integrations). Effectively, the front-end serves as the administrative user interface, allowing the owner of the PKG to manage their history (i.e., provide full read and write access) and synchronize with specific external services (which have been integrated by the service provider via their respective APIs). The application may be designed to work only online (i.e., live internet connection is required) or could allow offline access to the recommendations (but not to the PKG itself).

While this example solution seems feasible, it focuses on a single application and relies on tailor-made components. Having a PKG for a single application defies its purpose and offers limited benefits beyond increased transparency on the service provider's end. To unlock its potential, the PKG should amass data from multiple sources so that it could be utilized by multiple services. For example, there could come a time where data collected in the course of (uninjured) training using one application would be useful in the personalization of a rehab program provided by a different application. This requires standardization on the ecosystem level, that is, the use of shared vocabularies and communication protocols. Having a shared data representation is critical to ensure that users remain free to share their own data to a different service provider (e.g., different fitness application). We could envision a "PKG ready" badge system, similar to an ISO classification, for applications/services that meet this established standard. Note that this "PKG readiness" needs to extend to external services that allow for integration. For example, while Facebook and YouTube allow users to download an archive of their activity on the platform, there is no programmatic access to the same data via an API. This is not so much of a technical challenge, but rather a question of incentives. Established social media platforms have a strong (and from a business perspective quite understandable) motivation to keep users "locked in" within their own ecosystem. There need to be either financial incentives or regulatory frameworks to convince service providers to open up access to data via APIs using the PKG standards (and thereby help promote the PKG approach to taking ownership of user data utilized for personalized services). We discuss more about representation and management below, under Section 5.3.

In terms of the actual storage of PKGs, one possibility is to have them reside within accredited PKG hosting providers, similar to how other cloud-based services are hosted. This would provide users with a trusted and secure environment to store and manage their PKGs, following established standards for privacy and security. It would also

allow users to take their PKG to a different hosting provider, should they decide to do so. It is clear from Fig. 1 that the PKG and the management system around it are tightly coupled together, therefore the hosting provider would also need to provide an administration interface. Hosting providers could offer multiple options here, just like there are several options for virtual server management, including numerous open-source tools (cPanel, ISPManager, Webmin, etc.). The user-friendliness of this administration interface is of critical importance, as ordinary users need to be able to manage large amounts of data through this interface; see Section 5.4 below for further details.

It is worth pointing out that the PKG being under the owner's full control means that the owner can modify the PKG data. It is crucial that services take into account the possibility of incomplete or manipulated data when using PKGs. We illustrate this via our second example.

Example 2 (*continued*)**.** [Person] uses a PKG to manage their medical history, including records of treatments they have received in the past. They recently decided to share their PKG with a new doctor who would be treating them for a chronic condition. However, they were embarrassed about a particular medical procedure they had undergone in the past and did not want the new doctor to see it. Without much thought, they deleted the record of the procedure from their PKG. Unfortunately, the deleted record was relevant to the new doctor, who needed the information to properly diagnose and treat the patient's condition. The doctor's reliance on incomplete information from the PKG led to a misdiagnosis and ineffective treatment.

This example illustrates that the consequences could be severe and that "personalized service" must not be interpreted too liberally. Receiving recommendations about what specialist to visit given certain symptoms may be a service, provided that the user understands that (1) the responsibility of providing truthful and complete input to the service via the PKG lies with them and (2) the recommendations are not to be taken as medical advice, but suggestions for consideration. On the other hand, health care providers are responsible for educating patients about the importance of providing comprehensive and accurate medical history and relying on their own records, whenever possible. A potential solution is presented in Section 5.4.

In summary, we identify the establishment of PKG standards and their adoption by service providers as the most important open challenges for the realization of a PKG ecosystem, both of which would require a combination of incentive structures and regulatory control.

5.2. Population

Most research efforts relating to the aspect of PKG population concentrate on extracting structured data, like entities, relations, and attributes, from diverse data sources like tweets, academic articles, or food logs. However, the data extraction process usually restricts itself to a predetermined schema, e.g., a single vocabulary of predicates, which implies there is potential for further research. For example, how can a PKG population process discover novel predicates from unstructured data, assign these appropriately in the ontology, and subsequently apply these discovered predicates unambiguously? While data extraction from unstructured data is widely covered, only a few papers discuss the actual insertion of the extracted facts into the PKG, and even this coverage is quite light.

Example 1 (*continued*)**.** The Personal Trainer Assistant presents an immediate opportunity to use data collected by wearable electronics for personal fitness. However, a choice of schema or ontology must be made to have a shared vocabulary for integrating facts from disparate data sources such as the user's personal free text notes and data collected by wearable electronics from different manufacturers, which may not be all based on the same schema. The choices on this level also need to be informed by the specific intended uses of the application. For example, recommending a training plan depends on an ontology for characterizing exercises and their nominal properties. Based on different user goals (e.g., improving running form versus lifting heavier weights), the fitness application would match intended effects of exercises with user goals. The PKG population process may then include user-specific considerations, as different fitness goals imply different priorities about which facts to record in the PKG in order to track progress. It is likely most users would not wish to see these nuances surfaced after choosing a particular fitness goal, but rather it should be handled unobtrusively by the Personal Trainer Assistant. Note also that, since diet is an important component of personal fitness, while manually recording consumed calories is a challenge, it would be very useful to extend the knowledge extraction for PKG population to include facts extracted from images, e.g., inferring nutritional composition facts from a photo of a meal, which can then be used to log the user's diet.

In addition, the synchronization process has not been given enough attention, leaving room for future exploration on how to appropriately propagate changes made in the PKG to the private data sources. Synchronizing structured data with updates in the PKG is trivially automatable, but synchronizing may be more challenging in cases where the corresponding facts in private data sources are represented in unstructured form. Synchronizing unstructured data with updates in the PKG might need to rely on a notification system as proposed by Groza et al. (2007) and Sambra et al. (2016).

For example, a record of past events should generally not be changed, but a current prose description should be kept up-to-date. In this case, the fact to update may have been detected and extracted from one place in free text, but may also be expressed in different places in the text. These other places expressing that fact would then also need to be updated to fully synchronize the private data source with the PKG update. Identifying and tracking all such locations in unstructured data may present a challenge.

Example 2 (*continued*)**.** Sharing Health Information depends in large part on extracting facts correctly and unambiguously from free text clinical notes. Besides the challenges of discovering and incorporating novel entities and predicates from free text into the PKG, an important unsolved challenge for sharing health information is synchronizing updates in the PKG back to the private data sources. In clinical practice, the current state of a patient's case often needs to be reflected in a technical description. The records that constitute a patient's history may be where a fact was extracted and used to populate the patient's PKG, and these should generally not be modified, since the sequence of changes in a patient's history may be clinically important information. However, if the reality of the patient's condition changes and the PKG is updated accordingly, then a text such as the patient profile description intended to be read first by new clinical practitioners joining the treatment team must be up-to-date and reflect current facts. For a patient undergoing multiple treatments concurrently, a single patient profile may not be appropriate, as a dentist may need to see a different summary than an endocrinologist. Thus, coordinating all the places in unstructured data where a factual update in the PKG should be synchronized may be a non-trivial challenge.

From the two examples, we see challenges in both the process of populating a PKG with new extracted facts while conforming to a pre-determined target ontology, as well as the process of keeping selected external data sources synchronized with the current facts in the PKG. However, ubiquitous data capture through wearable electronics presents a great opportunity to apply the PKG concept in a useful and health-promoting manner.

5.3. Representation and management

While knowledge representation and knowledge base management are well established research areas, representation and management appears to be the least studied of the three identified aspects of PKGs. What seems to be particular for the representation of PKGs over regular KGs is the need for detailed contextual descriptions of the PKG facts, so that they can be correctly understood and interpreted at a later point in time and used for a wide range of purposes.

Another difference between PKGs and KGs is that while KGs are often centrally managed by a dedicated team, the responsibility of managing a PKG is ultimately its owner's, whose technical competence may be limited. There is therefore a strong need for tools that allow owners to manage their PKG efficiently without caring about the intricacies of data management and knowledge representation.

Example 1 (*continued*)**.** As the training preferences of the user are likely to be different depending on, e.g., the time of day, the time of the week, the time of year, and other personal events, such as traveling, the basic contextual data for all facts in the PKG should be recorded. By analyzing the user's data and context, the personal trainer assistant is able to provide more effective recommendations. For instance, it can identify that while the user typically favors indoor training during the winter months, an exception to this pattern is outdoor jogging in light rain. If training recommendations are given to the user based on these conclusions, it is important that an explanation for the recommendations is given which is rooted in the facts together with their context, so that the user can then better assess the given recommendations and adjust them as desired.

In this regard, a fruitful line of research would be to develop best practice modeling patterns for PKG statements that can gracefully cope with regular KG facts, provenance data, and statements about statements. This could take the form of a special purpose vocabulary designed to fit with a reification approach for statements, such as RDF's basic reification,[25] named graphs,[25] or RDF* (Hartig, 2017). A standardized vocabulary for PKG data would also be a huge benefit for the development of interoperable PKGs and PKG services. The PKG vocabulary and modeling patterns should accommodate expressing facts which may be stated using different, possibly incompatible, ontologies because the different facts in the PKG may be about unrelated domains. An early account of defining and using such a PKG vocabulary to translate natural language statements in a PKG via an easy-to-use web interface is demonstrated with promising results in recent work by Bernard et al. (2024).

A greater challenge seems to be that traditional reasoning techniques may not be directly applicable in such a scenario. On the one hand, since a PKG will typically make use of different ontologies and contain complex facts, such as temporal facts and facts of different modalities, we can assume that the required expressivity will go beyond the expressivity of OWL 2 DL.[26] OWL 2 DL is the most expressive fragment of OWL for which reasoning problems are decidable and reasoners are readily available, cf. Horrocks et al. (2006). As a consequence, standard OWL reasoning services may not be practically usable for most PKGs. On the other hand, the facts in a PKG may naturally be logically inconsistent, not because of modeling errors or poor data quality, but due to the diversity and "messiness" of real-world facts. Since standard reasoning services do not cope well (or at all) with logical inconsistencies, other techniques are needed.

Example 2 (*continued*)**.** The medical history of the PKG owner could be inconsistent for a number of reasons. For example, it could be because different doctors have made conflicting assessments of the owner's condition, or because the owner has manipulated the contents of the PKG. In such cases, the inconsistency should be identified and attempted to be fixed if possible, while also not disrupting reasoning services that operate on other, unrelated parts of the PKG.

To allow for reasoning over (parts of) the PKG data, both for discovering new facts and for maintenance tasks such as consistency checks, new strategies to ensure that reasoning is possible must be developed. One such strategy could be to carefully identify and isolate specific parts of the PKG for which reasoning may be performed using standard OWL reasoning techniques. A different strategy is to develop more fault-tolerant reasoners capable of handling inconsistencies, cf. Maier et al. (2013).

5.4. Utilization

The survey in Section 4 presents use cases for PKGs in diverse domains, mainly for the creation of personalized services. In our ecosystem, the same PKG is used by all the services, therefore, it represents resources with diverse formats (e.g., document, email, event, and text). One challenge for the service providers is to use the PKG data in an efficient manner. This includes seamlessly supporting this diversity of resources in addition to anticipating that some information can be manipulated or missing for several reasons such as access restrictions.

Example 1 (*continued*)**.** A PKG owner registers the following resources in their PKG: events attended, medical record documenting previous injuries, YouTube workout channels subscribed, and documents related to their diet. However, the owner decides to restrict access to its YouTube workout channels subscribed that can indicate how active the owner is and what type of exercise they prefer. Hence, the Personal Trainer Assistant should be able to suggest an adapted training plan despite the lack of information regarding the workout channels subscribed. To overcome this absence of information, the Personal Trainer Assistant uses reasoning over the owner's diet information to determine how active they are and uses this information as a substitute for the recommendation engine. Otherwise, the Personal Trainer Assistant can make a recommendation accompanied by a warning message indicating which missing information might improve the recommendation.

One solution proposed in this example is that service providers could utilize the accessible PKG data by inferring new knowledge about a user with reasoning techniques. However, using reasoning might not be that trivial due to different challenges such as the ones presented in the previous section. Therefore, the service providers could encourage the PKG owners to include truthful and complete information in their PKG and make them accessible via warning messages regarding missing information.

One long-term opportunity for PKGs lies in the administration process. Indeed, the survey in Section 4 shows that in general the owner of the PKG is not the user who receives personalized services, but the providers of these services. Consequently, the user does not have control over what is present in the PKG and who has access to it. Therefore, in the future, service providers ought to acknowledge that the ownership of the PKG belongs to the user and is operationalized via the administration process. The administration process serves two main purposes: the direct update of data inside the PKG and the control of its access by its owner. As mentioned before, allowing the PKG owner to directly interact with PKG data comes with potential risks such as the addition of untruthful information (e.g., wrong date for the last radiation session) and the deletion of critical information (e.g., removal of an allergy). This implies that the administration process should provide some guarantees and safeguards regarding data manipulation. One solution may be for service providers to require certain permissions with respect to the user's PKG in order to provide the service, as is current practice in many mobile applications. The access control needs

[25] http://www.w3.org/TR/rdf11-concepts/.
[26] https://www.w3.org/TR/owl2-overview/.

to consider different types of entities, i.e., a user, a group of users, and a service, that need read and/or write privileges. The existing tools (e.g., WebID, WebAccessOntology, and JSON Web Tokens[27]) in the field of semantic technologies might be used to build such process. Additionally, PKGs are not limited to a specific sample of the population, therefore, the administration process should be easy to use for everyone. An example would be to provide an intuitive user interface with menus and forms that abstracts the complex operations such as the creation and execution of SPARQL query to update the PKG. In recent work, Bernard et al. (2024) present a solution for populating and querying a PKG using natural language statements, which are automatically translated to API calls (and ultimately to SPARQL queries) by a backend service.

Example 2 (*continued*)**.** This example clearly illustrates the need for access control as it involves different actors (e.g., healthcare services and doctors). Indeed, one main question for the PKG owner is to determine what part of the PKG should be shared with whom. For example, they decide that the dentist does not need to know about the owner's last eye test in order to treat them. Therefore, they regulate access to the PKG using a group of external users/services per medical specialty. To do so, the interface of the administration process has a feature to associate external users/services to a specific group and can select the type of access (read/write), e.g., with a radio button. In terms of safeguards against data manipulation, medical service providers such as the dentist can require permission to keep their own unmodifiable copy of the PKG owner's medical history for the duration of the service provider relationship. Thus, if the PKG owner decides to remove critical information, e.g., removal of an allergy, from their PKG, the service providers can still make a reliable medical assessment.

To summarize, the administration process remains as a critical challenge for the adoption of PKG especially in the case of sensitive application domains where decisions can directly affect the PKG owner's life. The exploitation of PKG data also brings its share of questions to address, such as the treatment of missing or inaccessible data.

6. Conclusion

The aim of this study was to conduct a comprehensive survey and synthesis of the current research on personal knowledge graphs, with a focus on identifying critical gaps and key challenges that need to be addressed in order to enable the practical deployment of this technology. An important discovery from this paper is that there is no clear consensus on the meaning of a PKG. To address this lack of clarity, we have made explicit and compared the different ways in which the term is used, and proposed a new definition that emphasizes (1) data ownership by a single individual and (2) the delivery of personalized services as the primary purpose. Another key takeaway is the need to consider the larger ecosystem when conducting meaningful research on PKGs. We have proposed an unifying architecture and identified three main aspects that are required for successful development and deployment of PKGs: (1) population, (2) representation and management, and (3) utilization. Based on our survey of existing work, organized around these three key aspects, we make the following observations. First, it is apparent that certain aspects of PKGs have received more attention than others, and that a comprehensive, holistic approach is lacking. Second, different research communities (information retrieval, knowledge management, the Semantic Web, and natural language processing) can contribute to the development of PKGs in unique and complementary ways. Third, while specific building blocks exist, integrating them into a practical solution, even if only a rudimentary one, remains a difficult task, made even more challenging by the need for service providers to support and promote this technology. To drive adoption, service providers need to demonstrate the value that PKGs can bring to end-users. Finally, it is important to recognize that the biggest obstacles involved in developing and deploying PKGs are not of a technical nature, but lie in broader societal and organizational issues. To drive meaningful progress, it will be necessary for the research community, industry, and regulators to work together in a collaborative and strategic manner.

[27] https://jwt.io/introduction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the Norwegian Research Center for AI Innovation, NorwAI (Research Council of Norway, project number 309834).

References

Abrams, D., Baecker, R., Chignell, M., 1998. Information archiving with bookmarks: Personal web space construction and organization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '98, pp. 41–48.

Agichtein, E., Brill, E., Dumais, S., 2006. Improving web search ranking by incorporating user behavior information. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06, pp. 19–26.

Ammar, N., Bailey, J.E., Davis, R.L., Shaban-Nejad, A., et al., 2021. Using a personal health library–enabled mhealth recommender system for self-management of diabetes among underserved populations: Use case for knowledge graphs and linked data. JMIR Form. Res. 5 (3), e24738.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. Dbpedia: A nucleus for a web of open data. In: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. Springer, pp. 722–735.

Balog, K., 2018. Entity-Oriented Search. Springer Publishing Company, Incorporated.

Balog, K., Kenter, T., 2019. Personal knowledge graphs: A research agenda. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval. ICTIR '19, pp. 217–220.

Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., 2003. Taking email to task: The design and evaluation of a task management centered email tool. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '03, pp. 345–352.

Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., Whittaker, S., 2008. Improved search engines and navigation preference in personal information management. ACM Trans. Inf. Syst. 26 (4).

Bergman, O., Boardman, R., Gwizdka, J., Jones, W., 2004. Personal information management. In: CHI '04 Extended Abstracts on Human Factors in Computing Systems. In: CHI EA '04, pp. 1598–1599.

Bergman, O., Gradovitch, N., Bar-Ilan, J., Beyth-Marom, R., 2013a. Folder versus tag preference in personal information management. J. Am. Soc. Inf. Sci. Technol. 64 (10), 1995–2012.

Bergman, O., Tene-Rubinstein, M., Shalom, J., 2013b. The use of attention resources in navigation versus search. Pers. Ubiquitous Comput. 17 (3), 583–590.

Bernard, N., Kostric, I., Łajewska, W., Balog, K., Galuščáková, P., Setty, V., G.S., M., 2024. PKG API: A tool for personal knowledge graph management. arXiv:2402.07540.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. Sci. Am. 284 (5), 34–43.

Boardman, R., Sasse, M.A., 2004. "Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '04, pp. 583–590.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. SIGMOD '08, pp. 1247–1250.

Bush, V., 1945. As we may think. Atl. Mon. 176 (1), 641–649.

Capra, R., Khanova, J., Ramdeen, S., 2013. Work and personal e-mail use by university employees: PIM practices across domain boundaries. J. Am. Soc. Inf. Sci. Technol. 64 (5), 1029–1044.

Chakraborty, P., Dutta, S., Sanyal, D.K., 2022. Personal research knowledge graphs. In: Companion Proceedings of the Web Conference 2022. WWW '22, pp. 763–768.

Civan, A., Jones, W., Klasnja, P., Bruce, H., 2008. Better to organize personal information by folders or by tags?: The devil is in the details. In: Proceedings of the American Society for Information Science and Technology. ASIST '08, Vol. 45, pp. 1–13, (1).

Cui, L., Wei, F., Zhou, M., 2018. Neural open information extraction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). ACL '18, pp. 407–413.

Dou, Z., Song, R., Wen, J.-R., 2007. A large-scale evaluation and analysis of personalized search strategies. In: Proceedings of the 16th International Conference on World Wide Web. WWW '07, pp. 581–590.

Fader, A., Soderland, S., Etzioni, O., 2011. Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11, pp. 1535–1545.

Ferraiolo, D., Kuhn, R.D., Chandramouli, R., 2003. Role-Based Access Control. Artech House.

Fitchett, S., Cockburn, A., 2015. An empirical characterisation of file retrieval. Int. J. Hum.-Comput. Stud. 74 (C), 1–13.

Ganesan, B., Dasgupta, R., Parekh, A., Patel, H., Reinwald, B., 2020. A neural architecture for person ontology population. arXiv:2001.08013.

Gangemi, A., Nuzzolese, A.G., Presutti, V., Draicchio, F., Musetti, A., Ciancarini, P., 2012. Automatic typing of dbpedia entities. In: Proceedings of the 11th International Conference on the Semantic Web. ISWC '12, pp. 65–81.

Gerritse, E.J., Hasibi, F., de Vries, A.P., 2020. Bias in conversational search: The double-edged sword of the personalized knowledge graph. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. ICTIR '20, pp. 133–136.

Groza, T., Handschuh, S., Moeller, K., 2007. The NEPOMUK project-on the way to the social semantic desktop. In: International Conference on Semantic Technologies: I-Semantics 2007. In: I-Semantics '07.

Gurrin, C., Smeaton, A.F., Doherty, A.R., 2014. LifeLogging: Personal big data. Found. Trends Inf. Retr. 8 (1), 1–125.

Gyrard, A., Gaur, M., Shekarpour, S., Thirunarayan, K., Sheth, A., 2018. Personalized health knowledge graph. In: Joint Proceedings of the International Workshops on Contextualized Knowledge Graphs, and Semantic Statistics Co-Located with 17th International Semantic Web Conference. ISWC '18.

Hanrahan, B.V., Pérez-Quiñones, M.A., 2015. Lost in email: Pulling users down a path of interaction. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15, pp. 3981–3984.

Hartig, O., 2017. Rdf* and sparql*: An alternative approach to annotate statements in RDF. In: Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks Co-Located with 16th International Semantic Web Conference (ISWC 2017). In: CEUR Workshop Proceedings, Vol. 1963, CEUR-WS.org.

Heath, T., Bizer, C., 2011. Linked Data: Evolving the Web Into a Global Data Space. Morgan & Claypool Publishers.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Labra Gayo, J.E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J.F., Staab, S., Zimmermann, A., 2021. Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge, Morgan & Claypool, http://dx.doi.org/10.2200/S01125ED1V01Y202109DSK022, (22), URL https://kgbook.org/.

Horrocks, I., Kutz, O., Sattler, U., 2006. The even more irresistible SROIQ. In: Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning, KR'06. AAAI Press, Lake District, UK, pp. 57–67.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R., 2006. OntoNotes: The 90% solution. In: Proceedings of the Human Language Technology Conference of the NAACL. In: HLT-NAACL '06, pp. 57–60.

Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S., 2019. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. In: K-CAP '19, pp. 243–246.

Ji, H., Grishman, R., 2011. Knowledge base population: Successful approaches and challenges. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. HLT '11, pp. 1148–1158.

Jones, W., Bruce, H., 2005. Report on the NSF-Sponsored Workshop on Personal Information Management, Seattle, WA, 2005. Personal Information Management 2005: A Special Workshop Sponsored by the National Science Foundation.

Jones, W., Dinneen, J.D., Capra, R., Diekema, A.R., Pérez-Quiñones, M.A., 2017. Personal information management. In: Encyclopedia of Library and Information Science, Fourth Edition. CRC Press, http://dx.doi.org/10.1081/E-ELIS4-120053695.

Jones, W., Dumais, S., Bruce, H., 2002. Once found, what then? A study of "keeping" behaviors in the personal use of web information. In: Proceedings of the American Society for Information Science and Technology. ASIST '02, Vol. 39, pp. 391–402, (1).

Lin, T., Mausam, Etzioni, O., 2012. No noun phrase left behind: Detecting and typing unlinkable entities. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. In: EMNLP-CoNLL '12, pp. 893–903.

Lu, Y., Srivastava, M., Kramer, J., Elfardy, H., Kahn, A., Wang, S., Bhardwaj, V., 2019. Goal-oriented end-to-end conversational models with profile features in a real-world setting. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). In: NAACL '19, pp. 48–55.

Luo, L., Huang, W., Zeng, Q., Nie, Z., Sun, X., 2019. Learning personalized end-to-end goal-oriented dialog. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI '19, pp. 6794–6801.

Maier, F., Ma, Y., Hitzler, P., 2013. Paraconsistent OWL and related logics. Semant. Web 4, 395–427.

Matthijs, N., Radlinski, F., 2011. Personalizing web search using long term browsing history. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11, pp. 25–34.

Mazaré, P.-E., Humeau, S., Raison, M., Bordes, A., 2018. Training millions of personalized dialogue agents. In: The 2018 Conference on Empirical Methods in Natural Language Processing.

Nguyen, C.T., 2020. Echo chambers and epistemic bubbles. Episteme 17 (2), 141–161.

Panjwani, S., Shrivastava, N., Shukla, S., Jaiswal, S., 2013. Understanding the privacy-personalization dilemma for web search: A user perspective. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '13, pp. 3427–3430.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M., Vyas, V., 2009. Web-scale distributional similarity and entity set expansion. In: Proceeding of the 2009 Conference on Empirical Methods in Natural Language Processing. EMNLP '09, pp. 938–947.

Pasca, M., Durme, B.V., 2007. What you seek is what you get: Extraction of class attributes from query logs. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. IJCAI '07, pp. 2832–2837.

Rastogi, N., Zaki, M.J., 2020. Personal health knowledge graphs for patients. arXiv:2004.00071.

Sambra, A.V., Mansour, E., Hawke, S., Zereba, M., Greco, N., Ghanem, A., Zagidulin, D., Aboulnaga, A., Berners-Lee, T., 2016. Solid: a platform for decentralized social applications based on linked data. Tech. Rep., MIT CSAIL & Qatar Computing Research Institute.

Sarawagi, S., 2008. Information extraction. Found. Trends Databases (ISSN: 1931-7883) 1 (3), 261–377. http://dx.doi.org/10.1561/1900000003.

Seneviratne, O., Harris, J., Chen, C.-H., McGuinness, D.L., 2021. Personal health knowledge graph for clinically relevant diet recommendations. arXiv cs.HC/2110.10131.

Shen, X., Tan, B., Zhai, C., 2005. Implicit user modeling for personalized search. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. CIKM '05, pp. 824–831.

Shen, X., Tan, B., Zhai, C., 2007. Privacy protection in personalized search. SIGIR Forum 41 (1), 4–17.

Sontag, D., Collins-Thompson, K., Bennett, P.N., White, R.W., Dumais, S., Billerbeck, B., 2012. Probabilistic models for personalizing web search. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. WSDM '12, pp. 433–442.

Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R., 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '04, pp. 415–422.

Teevan, J., Dumais, S.T., Horvitz, E., 2005. Personalizing search via automated analysis of interests and activities. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '05, pp. 449–456.

Teevan, J., Dumais, S.T., Horvitz, E., 2010. Potential for personalization. ACM Trans. Comput. Hum. Interact. 17 (1), 4:1–4:31.

Teevan, J., Morris, M.R., Bush, S., 2009. Discovering and using groups to improve personalized search. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. WSDM '09, pp. 15–24.

Tigunova, A., Yates, A., Mirza, P., Weikum, G., 2019. Listening between the lines: Learning personal attributes from conversations. In: The World Wide Web Conference. pp. 1818–1828.

Tigunova, A., Yates, A., Mirza, P., Weikum, G., 2020. CHARM: Inferring personal attributes from conversations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 5391–5404.

Tiwari, S., Scharffe, F., Ortiz-Rodríguez, F., Gaur, M., 2023. Personal Knowledge Graphs (PKGs): Methodology, tools and applications. The Institution of Engineering and Technology.

Vannur, L.S., Ganesan, B., Nagalapatti, L., Patel, H., Tippeswamy, M.N., 2021. Data augmentation for fairness in personal knowledge base population. In: Trends and Applications in Knowledge Discovery and Data Mining. PAKDD '21, pp. 143–152.

Voit, K., Andrews, K., Slany, W., 2012. Tagging might not be slower than filing in folders. In: CHI '12 Extended Abstracts on Human Factors in Computing Systems. In: CHI EA '12, pp. 2063–2068.

Vrandečić, D., Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. Commun. ACM 57 (10), 78–85.

Weikum, G., Dong, L., Razniewski, S., Suchanek, F.M., 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. Found. Trends Databases 10 (2–4), 108–490.

White, R.W., 2016. Interactions with Search Systems. Cambridge University Press.
Whittaker, S., Matthews, T., Cerruti, J., Badenes, H., Tang, J., 2011. Am I wasting my time organizing email? A study of email refinding. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11, pp. 3449–3458.
Yang, T.-H., Huang, H.-H., Yen, A.-Z., Chen, H.-H., 2018. Transfer of frames from english FrameNet to construct Chinese FrameNet: A bilingual corpus-based approach. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). In: LREC '18.
Yen, A.-Z., Huang, H.-H., Chen, H.-H., 2019. Personal knowledge base construction from text-based lifelogs. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '19, pp. 185–194.
Zhang, S., Dinan, E., Urbanek, J., Szlam, A.D., Kiela, D., Weston, J., 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In: The 56th Annual Meeting of the Association for Computational Linguistics.