An approach based on Open Research Knowledge Graph for Knowledge Acquisition from scientific papers

Azanzi Jiomekong^a, Sanju Tiwari^b

^a University of Yaounde I, Faculty of Sciences, Cameroon fidel.jiomekong@facsciences-uy1.cm
 ^b Universidad Autonoma de Tamaulipas, Mexico, India tiwarisanju18@ieee.org

Abstract

A scientific paper can be divided into two major constructs which are Metadata and Full-body text. Metadata provides a brief overview of the paper while the Full-body text contains key-insights that can be valuable to fellow researchers. To retrieve metadata and key-insights from scientific papers, knowledge acquisition is a central activity. It consists of gathering, analyzing and organizing knowledge embedded in scientific papers in such a way that it can be used and reused whenever needed. Given the wealth of scientific literature, manual knowledge acquisition is a cumbersome task. Thus, computer-assisted and (semi-)automatic strategies are generally adopted. Our purpose in this research was two fold: curate Open Research Knowledge Graph (ORKG) with papers related to ontology learning and define an approach using ORKG as a computer-assisted tool to organize key-insights extracted from research papers. This approach was used to document the "epidemiological surveillance systems design and implementation" research problem and to prepare the related work of this paper. It is currently used to document "food information engineering", "Tabular data to Knowledge Graph Matching" and "Question Answering" research problems and "Neurosymbolic AI" domain.

Keywords: Digital libraries, Scientific papers, Open Research Knowledge Graph, Knowledge Acquisition, Knowledge management applications, Data and knowledge visualization

1. Introduction

Scientific papers are one of the greatest assets for scientists. They constitute one of the primary source of knowledge for researchers, and sometimes

for decision makers [1, 2]. They are recorded, indexed and disseminated in scientific publication repositories such as ISI Web of Knowledge, IEE Xplore, Springer, ACM, ScienceDirect, Scopus, Semantic Scholar, etc. In consequence, the body of scientific literature is growing at an enormous rate [2, 3, 4, 5]. This wealth of scientific knowledge is widely disseminated to users who now possess an unprecedented problem of access to scientific literature [4, 5, 6]. In effect, this increase in scientific content poses significant challenges for the researchers who want to sort through, read, understand, compare, and build upon to determine for instance, the state of art in their respective field of interest [4, 7].

Globally, a scientific paper can be divided into two major constructs which are Metadata and Full-body text [4, 5]. Metadata provides a brief overview of the scientific papers and the Full-body text contains valuable information that is beneficial to fellow researchers. To retrieve metadata and key-insights from scientific papers, Knowledge Acquisition (KA) [8] is a central activity in research.

Knowledge are facts, information and skills acquired through experience or education for understanding of a subject area [9]. Concerning scientific papers, knowledge are metadata provided by editors and authors, and key-insights provided by authors which are used by fellow researchers to understand the scientific paper content. Knowledge Acquisition from scientific papers refers to the method for gathering, analyzing and organizing knowledge embedded in these papers. This involves the extraction of structured content in the form of entities, relations, facts, terms, and other types of information that may help researchers to understand the papers and get insights from them [6]. After its acquisition, knowledge is organized in such a way that it can be used and reused whenever needed. Globally, knowledge acquisition can happen through a wide variety of strategies that vary from completely manual to totally automated [1, 7, 8]. Concerning knowledge acquisition from scientific papers, we distinguished the manual process [1, 7, 10, 11] and the (semi-)automatic process.

Given the amount of scientific papers that a domain may have, the manual process can be a cumbersome job, time consuming, not scalable and not efficient. To reduce the burden of KA, computer-assisted and (semi-)automated strategies are proposed [2] for processing and cataloging scientific knowledge, for assisting researchers to choose their papers, navigate amongst papers, compare them and get insights from them.

During the last decades, many researchers have contributed to the automatic extraction of metadata from scientific papers. Multiple rule-based, machine learning and NLP techniques have been proposed [1, 4, 5]. Con-

cerning knowledge extraction from the full-body text, it has been reported that key-insights are deeply hidden in the text and are difficult to extract [1, 2, 3, 4, 12]. To allow researchers to collaboratively build the body of knowledge from their domain and research interest, we propose a computer-assisted knowledge acquisition approach. This is based on the use of Open Research Knowledge Graph (ORKG) [3] for automatic acquisition of meta-data and manual annotation of the paper with key-insights to produce a semantic description of the scientific knowledge of the domain in a Knowledge Graph (KG). Once extracted and organized, research contributions can be compared using annotated tables and graphics.

This approach is inspired by the use of ORKG in our research since three years to: (1) Organize and compare research contributions so as to build a large dataset of up-to-date knowledge for the following research problems: "ontology learning", "epidemiological surveillance systems design and implementation", "food information engineering", "Tabular data to Knowledge Graph Matching", "Question Answering", and "information extraction from scientific papers", and "Neuro-symbolic AI" domain. (2) Organize research so as to facilitate the update and improvement with the contributions of fellow researchers working on the same research problem or the same domain.

In the rest of the paper, we present Open Research Knowledge Graph in Section 3 and the research methodology in Section 4. In Section 5, we present the approach we propose for Knowledge Acquisition from scientific papers using ORKG and in Section 5 we present the use of this approach on 3 use cases: "epidemiological surveillance systems", "food information engineering" and "knowledge extraction from scientific papers". The latter use case was used to write the related work of this paper (Section 7). Finally, in Section 8, we conclude.

2. Scientific papers description

On the basis of its structure, knowledge contained in a scientific paper is broadly classified into two major categories which are metadata (see Section 2.1) and key-insights (see Section 2.2).

2.1. Metadata

Metadata information is used either for scientific paper recommendation by research repositories, or to furnish a brief overview about a scientific paper. The latter allows a researcher to decide the paper's relevance with their domain of interest [4]. Metadata can be defined in two main components: those that are assigned by the authors (such as the title of the paper, Abstract, Keywords, etc.) and those that are assigned by the editors (such as BibTex and/or DOI, Copyright, Date of publication, etc.).

Metadata extraction. Metadata extraction (ME) refers to the identification and extraction of metadata elements. In order to perform ME, there exist multiple datasets that vary on the basis of article's sources, publication venues, etc. On these datasets, multiple automatic approaches are applied. They use the DOI, BibTex or the title of the paper to search and fetch these papers from scientific repositories. Thereafter, rule and/or Machine Learning techniques are used to extract these metadata [4].

2.2. Key-insights

The full body text of the paper hides the key-insights/knowledge that the readers need to extract in order to understand the paper. Even if the authors can choose their own way to organize the full-body text, the journal's Guide for Authors provides to the authors a template composed of the different sections that the paper may include. Whatever the organization of the paper provided by the authors, one can identify the introduction, Research methods and methodologies, Results, Discussion, Related work and/or literature review and conclusion.

Knowledge extraction from the full-content of a scientific paper. From the full-body text of a scientific paper, entities such as research domain, research problem, methodology of the research, methods, models, algorithms, processes, data-source, data-sets, tools, evaluation measures, results achieved, limitations of the research, future directions, etc. are extracted by the readers in order to understand the paper. These entities once extracted can be organized into instances. These instances can be grouped into classes with associated properties.

From classes, the following relations can be extracted:

• Taxonomy: this relation organizes classes into a hierarchical relation. For instance, we used it to organize ontology learning research problems using a taxonomy of research problems related to ontology learning. This taxonomy shows that ontology learning research can be divided into the following research problems: "Ontology learning from unstructured sources", "Ontology learning from semi-structured sources", and "ontology learning from structured sources". These research problems can also be divided by considering different data sources.

• Association: This is the link used to define that two classes are related to each other. For example, in the sentence: "Jiomekong et al. proposed to use Hidden Markov Models to extract knowledge from source code", we identified the classes "Techniques" and "Knowledge source". So, a relation named "extract" of application between the class "Technique" and the class "Knowledge source" can be established. The instance of the class "Technique" is then "Hidden Markov Models" and the instance of the class "Knowledge source" is "Source code" and we can have the following statement: "Hidden Markov Models are used to extract knowledge from source code".

Once extracted, key-insights are grouped into research contributions and used to write state-of-the-art. In the latter, many tables and graphics are used to compare research contributions of several authors.

The extraction of key-insights from scientific papers are generally manual. However, knowledge extracted are sparse in different data sources (scientific papers, research computers, etc.), with the risk of being forgotten, lost and making it difficult to compare background research problems to upto-date ones. In the next section, we present how Open Research Knowledge Graph can be used as a computer-assisted tool to solve these problems.

3. Open Research Knowledge Graph

In this section, we present an overview of ORKG (Section 3.1) and the main features used during this research (Section 3.2).

3.1. Overview of ORKG

ORKG is an open research infrastructure designed to acquire, publish and process structured scholarly knowledge published in the scholarly literature [3, 13]. It is built according to the principles of Open Science, Open Data, and Open Source.

- Open Science: ORKG resources such as comparisons of scientific papers and the smart reviews can be developed through collaborative network of researchers. Once published, these resources are freely available to anyone who wants to learn about the research question and/or to contribute.
- Open Data: All the data ingested in ORKG is in a machine readable format and open to everyone who needs to share, use, re-use, modify,

and share the modified version. The only restriction concerns contributing to an ORKG resource. This restriction consists of having an ORKG account.

• Open Source: the source code of ORKG is available to the general public.

Thus, all the ORKG source code, information, data are available under open licenses [3]. To date, ORKG indexes more than 10,000 research papers corresponding to more than 5000 research problems (corresponding to 1237 research fields), more 1000 comparisons, 224 templates, 1000 users, 2216 benchmarks¹.

3.2. ORKG features

To help researchers structure and organize the research contributions extracted from scientific papers, ORKG provides a set of features. In this Section, we present the ones we used during our research.

Add research problems. The research problems of a research area can be described independently, provided with relevant sources and assigned to a taxonomy of research problems [3]. For instance, with ORKG, we can define a taxonomy of research problems related to ontology learning.

Add papers. ORKG represents an article with [3]:

- 1. **Article metadata:** The article metadata involves the bibliographic information such as article title, authors, journal, book title, etc.;
- 2. Semantic description of the article: These are key-insights of the papers extracted and annotated by researchers by following the Subject-Predicate-Object triple principle.

The article metadata and its semantic description are used to annotate the paper. To this end, researchers are allowed to add papers manually or (semi-)automatically to ORKG (see Fig. 1) [3]:

• During the manual process, all the key-metadata (title, author, etc.) and key-insights (research domain, research problem, research tools, etc.) of the papers are manually acquired by the researchers and added to ORKG using a wizard provided by the system.

https://www.orkg.org/orkg/stats

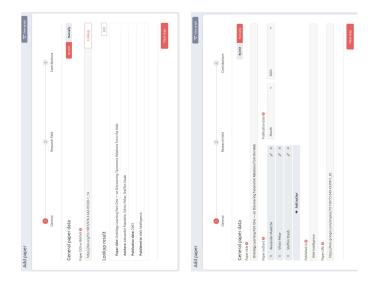


Figure 1: Manual (first picture) and automatic (second picture) acquisition of meta-data of a paper

• To semi-automatically add an article to the system, the key-metadata of the article such as the paper title, DOI or BibTex are provided to the ORKG wizard. These informations are used by the system to fetch the articles key-metadata. Once extracted, these informations are presented to the users so that they can complete missing metadata. Once the metadata are added to the paper, the researchers use a wizard provided by ORKG to semantically describe the paper with key-insights they extracted manually.

Semantic description of research papers. The semantic description of research papers consists of the annotation of these papers with key-insights extracted from them and to organize these elements into research contributions. This allows us to put the paper in machine readable form following the RDF subject-predicate-object paradigm. The ORKG annotation feature is a flexible system that allows users to reuse existing predicates and functions or to create and add their own predicates (properties or attributes). The description of the entities in human readable form allows researchers to have a common understanding of the data between the various stakeholders. Fig. 2 presents an example of a graph representation of a paper with their metadata and key-insights organized in paper contributions.

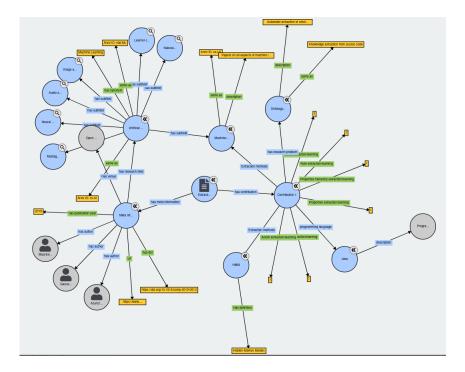


Figure 2: RDF graph representation of a paper with their metadata and contributions

The graph of Fig. 2 presents the paper entitled "Knowledge extraction from java source code using Hidden-Markov Models". The key-insights extracted are:

- The research problem which is "Knowledge extraction from source code",
- Different types of knowledge that are extracted,
- Techniques that are used during the extraction process,
- The programming language from which the source code was written.

Add research contributions to a paper. In ORKG, each paper consists of at least one research contribution which addresses at least one research problem and is further described with contribution data including materials, methods, implementation, results or other key-insights. The paper of the Fig. 2 presents one research contribution. These contributions can be compared between them or by other contributions from other papers [13] in an ORKG



Figure 3: Adding a paper using the comparison table wizard

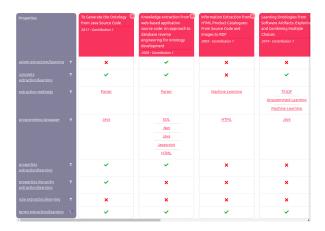


Figure 4: A table presenting the comparison of research contributions of papers related to ontology learning from source code

comparison table. Papers can be added to ORKG during the creation of the comparison table as presented by the Fig. 3.

We consider in this research that all the key-insights in the research paper such as the definition of research problem, materials and methods used, results obtained, lessons learned, etc. are grouped into research contributions. During the adding paper process, a default research contribution containing the key-insights such as research domain and research problem is filled by the user. Research contributions are described in a structured and semantic way as a Knowledge Graph (see Fig. 2). Therefore, the information will not be only readable by humans, but also by machines [3].

Comparing research papers. The structured content descriptions of scientific contributions presented above are presented in such a way that the contribution becomes comparable with other articles of the research domain. There-

fore, the structured semantic representation of scientific knowledge in the knowledge graph makes it possible to automatically create literature comparisons. Allard et al. [13] present a workflow designed to compare research contributions in ORKG. Fig. 4 presents an example of a comparison table built using this workflow. This is the comparison of research contributions of papers related to ontology learning from source code. The comparison table can be published with DOI, exported in various formats such as RDF, LaTeX, PDF, CSV and integrated in a literature review. The comparison table link can be shared to other researchers, so that they can improve the comparison by correcting errors or adding missing information.

Templates. Scientific papers usually lack a formal metrical structure. It comprises full grammatical sentences, paragraphs in which key-insights are hidden. Identifying and structuring research contributions found in scientific papers is not always an easy task for a research student or newcomers in the domain. This is because the description of scientific findings is complex and is based on expert knowledge. On the other hand, the researcher should decide in which granularity a research contribution should be described so as to be comparable.

The goal of the template is to highlight for a research problem, a set of key-insights that may be found in a scientific paper addressing this research problem. It specifies the structure of scientific information so that [3]: (1) Fellow researchers can compete with more key-insights, (2) New researchers can rapidly get insights in the research domain.

Templates can then be reused in the description of research contributions to facilitate data entry and ensure comparability. For instance, we built a template for documenting existing datasets for metadata extraction from scientific papers ².

Graph visualization. Once added to a paper, the graph representing the research contribution is generated. This graph can be used for the exploration of scientific contribution.

Importing survey papers: . Survey articles present an overview of the state-of-the-art for a specific area. Within survey articles, some overviews or summaries are often presented in (semi-)structured tabular format. From these tables, information on key-insights of the papers involved in the literature review can be extracted (semi-)automatically as follow: the first step

²https://www.orkg.org/orkg/template/R277000

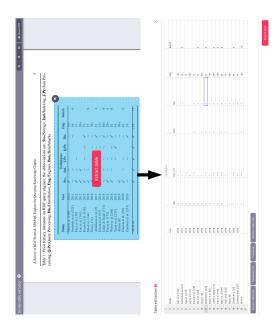


Figure 5: Extracting key-insights on graph databases using ORKG

consists of extracting the key-metadata and the key-insights from the table and building a comparison table; the second step involves fixing potential extraction errors and adding additional metadata or key-insights that was not automatically extracted. The Fig. 5 presents the extraction wizard.

Smart Review. After the creation of a comparison, a researcher may create a smart review for giving an overview on research addressing a particular research question. To this end, ORKG furnishes a "What You See Is What You Get (WYSIWIG)" editor allowing researchers to create a structured overview of the literature.

Collaborative work on literature review. In ORKG, collaborative work allows a whole community of researchers to collaboratively build the state of the art of a research problem. In effect, many authors working on the same research problems can gather to add and modify research contributions of a scientific paper. Once these contributions are compared using ORKG comparison tables and used to write smart reviews they can be shared amongst other researchers in order to get their viewpoints. To this end, contributions and smart reviews are versioned so that all changes can be discussed by the

professional community, updated and new versions published. If new literature is published, it is easy to continuously expand the comparison, which thus continues to reflect the current state of knowledge in a comparable way.

4. Research Methodology

The research methodology consists of action research. Action research methodology is used when major challenges cannot be studied without implementing them, and where implementation implies a long term commitment because effect may take time to emerge [14]. In our case, we wanted to explore, test and evaluate the different features that can be used for knowledge acquisition from scientific papers using Open Research Knowledge Graph as a computer assistant tool. Given that action research allows us to plan, implement, revise, then implement, lending itself to an ongoing process of reflection and revision, we thought it necessary to use this research methodology.

Globally, the research methodology consisted of a set of aggregated interventions to curate the papers. These interventions involved a series of actions taken during the curation of scientific papers. At its end, we come up with a research methodology that is reported in this section. This research methodology consists of the Pre-Intervention (Section 4.1) and the Intervention (Section 4.2) phase of the Action research methodology. The Post-Intervention presented in Section 6 consists of the use of the methodology presented in this section in three use cases.

4.1. Pre-Intervention

During the ORKG curation, we particularly worked in the domain of Semantic Web. The Pre-Intervention step consists of the definition of the research objective and the organization of the curation.

4.1.1. Research objective

We started this research in 2021 with the objective to document the research problem: "ontology learning". In effect, ontology learning is the automatic or semi-automatic extraction of ontological knowledge from unstructured, semi-structured or fully structured knowledge sources in order to build an ontology from them with little human intervention [15, 16, 17, 18]. This choice was motivated with our recent work on ontology learning from source code [15]. This is the automatic extraction of ontological knowledge from software source code. Thus, we decided to curate this paper first. Thereafter, we curated the related work of this paper.

4.1.2. Selection of papers to curate

We started with the selection of the paper we wrote on ontology learning from source code [15]. Thereafter, we selected all the papers related to ontology learning from other data sources that were cited in this paper. For the "ontology learning from datasources" that were not cited such as "ontology learning from folksonomies" or "ontology learning from thesaurus", we used the famous research repository Semantic Scholar to search for relevant papers. The keyword "ontology learning from xxx" (where xxx represents the data source) were entered in the search bar of the Semantic Scholar platform. We used the papers titles, short abstract provided by Semantic Scholar and paper abstract provided by the authors to select relevant papers. Given that our goal was mainly to curate some papers and understand how to use ORKG for knowledge acquisition from scientific papers, we only choose the papers on the first page results. Key-insights were extracted from these papers, comparison metrics defined and used to compare these papers.

4.1.3. Work organization

Globally, the curation of ORKG involves two groups of people: the ORKG team and the curators. The ORKG team is a group of persons responsible for the organization of the curation meetings, description of tasks of curators, training of curators on the use of the tool and support when they have any difficulties. Before we start the curation in June 2021, a training session was made by the ORKG team. This session was oriented on the presentation of ORKG features, the creation of comparisons using the ORKG comparison editor. During the period of curation, many demos on the creation of comparisons, templates, and smart reviews were made.

To support the curators and respond to all their difficulties, a Mail and a Skype Group were created and a bi-monthly meeting was set-up. During these meetings, we were having 5-10 minutes time to present our work: adding papers, creating comparisons tables, templates, smart review, etc. Thereafter, the questions and the remarks were posed in order to help to ameliorate the work. The meetings were recorded with Skype so that we can watch it later. During these meetings, the comparisons of papers made by the curators were discussed so that they can update and correct errors. Examples of discussions concern the definition of classes, properties, the coding of knowledge extracted from the scientific papers, etc.

4.2. Intervention

During the intervention phase, we extracted key-insights from scientific papers and we used these key-insights to create comparison criteria (these are ORKG properties). Thereafter, these comparison criteria were used to compare these scientific papers using the ORKG comparison table. This is an iterative and incremental process during which the experience we got during the creation of one comparison table is used to ameliorate this comparison table and create the new ones. Comparison tables were evaluated by the ORKG team and fellow researchers and refined. For instance, the first comparison³ were refined until it was accepted as well organized by the ORKG team and some colleagues working in the domain of ontology learning. Globally, papers related to the following thematic were curated:

- Ontology learning from Thesaurus (5 papers),
- Ontology learning from Glossaries (2 papers),
- Ontology learning from taxonomies (2 papers),
- Ontology learning from XML (15 papers),
- Ontology learning from UML (4 papers),
- Ontology learning from source code (9 papers)
- Ontology learning from folksonomies (6 papers)
- Ontology learning from images (2 papers)
- Ontology Learning from Entity Relation Model (9 papers).

At the end, 54 papers were curated, 9 comparison tables were created using these papers, and one smart review on ontology learning from images. In the following paragraphs, we present how we proceed to create these comparisons, the lessons learned and main finding that was used to ameliorate our work.

4.2.1. Creation of the first comparison

The first work we did was to create the first comparison of papers. Nine papers related to "ontology learning from source code" research problem were read, knowledge extracted and ingested into the ORKG platform. To this end, we firstly created a comparison table and using the ORKG comparison table wizard, we added papers to ORKG. These papers were added manually and (semi-)automatically to ORKG:

³https://www.orkg.org/orkg/comparison/R138057

- We used the manual process for the papers that do not have DOI or BibTex. During this process, all the key-metadata (title, author, etc.) and key-insights (research domain, research problem, etc.) of the papers are manually acquired and added to ORKG using a wizard provided by the system.
- To (semi-)automatically add an article to ORKG, we used the DOI or BibTex to automatically fetch the articles metadata. Once extracted, missing informations are completed and the paper is annotated with key-insights extracted manually.

Once a paper is added, a graph representing the research contribution allows us to visualize and verify that the information on the paper is well structured.

The comparison table of ontology learning papers from source code contains the following elements:

- The first column of the table contains properties, which can also be seen as a comparison criteria.
- The rest of the column corresponds to papers that are compared.
- For each row, the corresponding insight extracted from the paper is presented, so that these elements can be used to compare papers together.

From this comparison, we learned how to organize research contributions using ORKG. The exchange with the ORKG team and some colleagues working in the domain of knowledge engineering allowed us to ameliorate this comparison and a new version was published. We found the tool interesting to save our work so to reuse later in scientific papers as additional material or related work. This motivates us to create more comparisons and explore the other features of the system.

4.2.2. Creation of other comparisons

The creation of the first comparison allowed us to master the use of the comparison wizard. Therefore, 7 more comparisons were created. These comparisons gave rise to a refinement iteration in order to identify all potential knowledge that will be converted into classes, relations and properties and that will be used to build a high-quality and comparable structured scientific knowledge for "ontology learning" research problems. The aim of this structure being to create a common Semantic Model to reflect contributions

to "ontology learning" research problems. For instance, for ontology learning methods such as "TF.IDF", "Unsupervised Learning", "deep learning", "Neural Network", we decided to group them and to create a class labeled "Learning method".

Lesson learned. The comparisons presented above led to the following lessons:

- Structure and describe research contribution is not an easy task: During the creation of comparisons presented above, we learned that to structure and describe a research paper is not an easy task. In effect, describing research contributions and making them comparable is complex because the granularity of comparison should be decided. For instance, should we consider the comparison of methods for knowledge extraction from "unstructured sources" and "structured sources" or should we go further and compare unstructured data sources such as "text", "images", with the structured ones such as "databases", "UML models"? Given that we wanted fellow researchers to see the methodologies, methods and tools for ontological knowledge extraction from knowledge sources, we decided to add a property that indicate if the data source is unstructured and the type of the data source (e.g., "text", database", etc.)
- Find the accurate property for the comparison is not an easy task: It is recommended to reused as much as possible existing ORKG properties that were created by other researchers. However, we found it difficult because one has to scroll down any time one wants to add a property to a contribution (time consuming). On the other hand, after some time, the description of a property can be forgotten or unknown (for those who did not input them). This makes it difficult to find the right property to use in the comparison tables. Fortunately, the ORKG wizard provides the properties description. However, many properties had the same name and no description.

Insight. To solve the above problems, we found it necessary to use the ORKG template feature to structure scientific papers related to "ontology learning". This template is supposed to contain all the properties that should be compared. To facilitate its accessibility, we decided to add descriptions to all the properties used. Thus, to add a contribution from a paper related to the "ontology learning" research problem, this template is used. This template is a standardized tool that can be refined and used to compare as many scientific papers of this research problem. The creation and the use of this template is presented in the following paragraphs.

Table 1: Table describing the classes of the template used to describe contributions of papers related to ontology learning

Class label	Example of instances
Knowledge source	Text, databases, source code, etc.
Learning purpose	Constructing a new ontology, updating an existing ontology
Application domain	Medicine, Geography
Learning data source	Java source code, XSD documents
Has dataset	300 source code files selected in the data source
Training corpus	70% of the dataset
Output format	.txt, .owl, .json, .rdf, .xml
Input format	.txt, .XML, .png
Learning method	Parser-based, Machine Learning-based, HMM, CNN
Learning tool	on-to-text, source2onto
Technologies	Java, Python, TensorFlow
Terms learning	Entities, shape, feature, aspects
Relationship	Topological relation, Direction relation
Property	DataProperties, ObjectProperties
Axiom	Transitive relation, reflexive relation
Rule	if(age;10)then children
Evaluation	User evaluation, comparison to a gold standard
Knowledge assessment	Empirical measure, human intervention, domain expert

4.2.3. Template creation

After many comparisons, we found it necessary to provide a structure to organize the knowledge extracted from papers related to ontology learning. This structure allowed us to facilitate the organization of further relevant papers independently of the curator in a highly consistent knowledge graph.

To create the template, we used the properties we already added in the system for ontology learning from source code, database, UML models, etc. This template involves classes, properties (presented by the tables 1 and 2) applicable to a considerable number of papers related to ontology learning research papers. The comparisons elements that are created using this template are composed of instances of these classes and relations included in the template.

Each class is associated with a property that will appear in the comparison table as a comparison criteria in column property of the comparison table. In addition to these properties, other properties of basic data types are also added to the template. These properties are presented in the table 2.

Table 2: Properties for comparing research contribution

Property label	Description
Class learning	True when the authors extracts classes from the data source
Instance learning	True when the authors extracts instances from the data
	source
Taxonomy learning	True when the authors extracts taxonomies of classes or
	properties from the data source
Class hierarchy learning	True when the authors extracts class hierarchies from the
	data source
Validation tool	Presents the technologies used to validate/develop the vali-
	dation tool
Validation comments	Any comments of the authors concerning the validation
Recall	This is the recall of the learning tool
Precision	This is the precision of the learning tool
F-measure	This is the F-measure of the learning tool

4.2.4. Using the template to create a new comparison

The template presented in the section above was used to create 14 contributions. These contributions come from 2 papers related to ontology learning from images. To create these contributions, we identified the DOI of the papers found using Semantic Scholar. The DOI was entered using the "adding paper wizard" of ORKG. The system automatically extracts the papers metadata. Thereafter, knowledge was extracted manually and added to the system using the template. These contributions were finally used to create a comparison table. The graph visualization was used for the exploration of scientific contributions. It allowed us to realize that there was some confusion in our comparison. This confusion was corrected, the template and the comparison refined and new versions published. A video presenting the curation of papers related to ontology learning from image was published by the ORKG team⁴.

4.2.5. Creation of smart review

Once the informations are extracted from the papers related to ontology learning from images, these information were used to write a smart review. The goal of this review was to present and compare related work on ontology learning from image data.

 $^{^4 {\}tt https://www.youtube.com/watch?v=EwfLJdPRr6o}$

4.2.6. Collaborative work on literature review

In this research we did not consider only our viewpoint during the creation of the template and comparison tables. We discussed with colleagues, other researchers using ORKG and the ORKG team to which we sent the links of these resources. This allowed us to refine them and create new versions. It should be noted that any fellow researchers can improve these resources with new information. For instance, if new literature is published, anyone can add a new contribution to the comparison table and publish a new version.

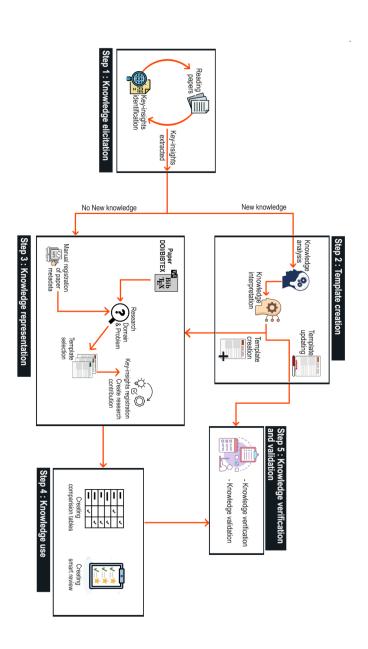
5. An approach for knowledge acquisition from scientific papers

Acquiring knowledge from scientific papers from scratch is costly in time and resources. The approach we propose in this paper aims to reduce this cost during the knowledge acquisition process by allowing researchers to create structured repositories of scientific papers related to a research problem and/or a research domain. This approach is inspired by the use of ORKG in our research since three years to:

- Organize and compare research contributions so as to build a large dataset of prior and up-to-date knowledge in our research domain;
- Organize research so as to facilitate the update and improvement with the contributions of fellow researchers working on the same research problem.

In effect, previously, to do a state of the art research, we were searching for relevant scientific papers on the Internet, reading these papers and summarizing them in text format and building comparison tables using LibreOffice Calc and Google Sheet. After the curation of ORKG in 2021, we got new insights on how to acquire and organize scientific literature. The latter is developed in this section as a computer-assisted knowledge acquisition approach from scientific papers (presented by Fig. 6). It describes how knowledge can be extracted from research papers and stored in a knowledge graph in order to facilitate the access to key-insights hidden in research papers. It consists of six steps during which classes, properties and relations are extracted from scientific papers, and used to build a template. Thereafter, the template is used to represent contributions of the papers related to the same research problem. Finally, the contributions are used to build comparison tables, which themselves can be used to write a smart review. These steps are: Knowledge elicitation (Section 5.1), Knowledge

Figure 6: The description of the knowledge acquisition approach proposed in this paper



analysis and interpretation (Section 5.2), Templates creation (Section 5.3), Knowledge representation (Section 5.4), Knowledge use (Section 5.5) and Knowledge verification and validation (Section 5.6).

5.1. Step 1: Knowledge elicitation

First and foremost, the researcher should determine the research domain that he wants to document. Thereafter he should identify the research problem related to this research domain. Once the research domain and the research problem are identified, these informations are used to search for relevant papers using search engines like Google search or search engines in digital research repositories like Semantic Scholar, Springer, Elsevier, IEEE, etc. For instance, in the domain of nutrition, a researcher may be interested in the food recommendation to people on diet. Thus, the following research question may be elicited: "How to recommend food to people in diet?" or "which techniques, methods and methodologies are used for food recommendation?". These research questions are used to search for scientific papers. Relevant papers related to this research domain and research problem are identified using many criteria which can be the title of the paper, the authors, references or citation analysis. References analysis can be used for instance to identify relevant papers to the research problem. During the selection of papers, the importance of a paper is defined as how close it is with the research domain and research problem. This task is done by reading the abstract or the full paper. Only papers that are too close to the research problem are selected. Once the research papers are found, some of them are selected for knowledge elicitation.

During the knowledge acquisition activity, the researcher should read the papers selected previously, identify and extract keywords, clauses, sentences, scientific claims, etc. Globally, all the information that is relevant to understand the paper is identified. This is an iterative process (see Fig. 6) during which the researcher should be sure at the end that he has identified anything relevant. In early iterations of the cycle, the knowledge identified can refer to entities which are grouped, and will give classes. These classes will thus be put in relation with each other. At the end of this step all the relevant knowledge are extracted.

The identification and the extraction process can be done by using hand-written notes, spreadsheet, or underline in order to highlight all the key-insights. Thereafter, each piece of information highlighted can be labeled with the type of knowledge it represents. For instance, if we highlighted "HMM is used to extract information from source code", then, we can label "HMM" as a Machine Learning method, "Source code" as a knowledge

Figure 7: Representation of the triple: "HMM is used to extract information from source code" $^{\prime\prime}$



source and "extract information from" as a relation between the ML method and the knowledge source (Fig. 7 presents this triple).

Given that within survey papers some overviews or summaries are often presented in (semi-)structured tabular format, the comparison criterion in these tables should be identified and extracted. These information could be extended with additional information extracted from this survey paper or other papers selected.

Globally, two kinds of information can be identified from the papers. We named them as keywords and keyphrases:

- **Keywords:** keywords are words that are used to represent knowledge. For instance, if we consider the evaluation of ML techniques, we can identify the following keywords: "HMM", "Recall", "Precision", "Accuracy".
- **Keyphrases:** keyphrases are composed of a set of words that are used to represent a part of knowledge. For example, we have "Source code", "Wind power forecasting using time series".

5.2. Step 2: Knowledge analysis and interpretation

Knowledge analysis and interpretation consists of reviewing the elements extracted, identifying the key pieces of knowledge, providing a definition to each of these elements. Thereafter, these knowledge are assembled into related groups. Redundant informations are identified and only one term is selected. The definition of each keyword and keyphrases is provided.

Knowledge obtained after this task are classified into classes, relations, properties and instances. The terms in keywords and keyphrases are used to create the labels of these entities. During this task, the main challenge is to keep the keywords and keyphrases simple and descriptive.

5.3. Step 3: Template creation

The classes, properties, relations and instances are used to create a template using ORKG template editor. This template is a conceptual model of papers dealing with the research domain and research problem addressed by its creator.

The template allows researchers to put key-insights hidden in research papers in a machine readable form. However, to be human readable, Classes, relations, properties and instances should have a definition in human readable form so that any human operator can use the template to register knowledge extracted from a paper. In order to create a consensus, the template link can be sent to researchers working in the research domain to have their point of view. To facilitate her/his amelioration, the author of the template can make the latter editable, so that other researchers can update.

5.4. Step 4: Knowledge representation

The knowledge representation step consists of using the template built in step 3 to annotate research papers related to the research domain and the research problem. Thus, the research contribution is machine and human-readable. Using the template, knowledge related to the research problem and research domain is continually refined and updated through additional knowledge from new scientific papers.

Globally, annotating a paper using ORKG and the template built in step 3 can be manual or (semi-)automatic. During the automatic process, the paper title, DOI or the BibText is entered in the add paper wizard. These metadata are used to fetch the paper and automatically extract other metadata. The next step of the process consists of selecting the research domain, defining the research problem and choosing the template to use in order to fill the other key-insights. Importing survey tables are also done (semi-automatically). Once the table is imported, the curator can correct information extracted and add additional key-insights. The manual process consists of adding the metadata and the key-insights manually.

Once ingested into ORKG, research contributions can be visualized as a semantic network. This graph can be used for the exploration of scientific contributions.

5.5. Step 5: Knowledge use

Extracting knowledge from knowledge sources is not an end in itself. Once represented in a machine readable form, the knowledge acquired should be used. In our case, the knowledge acquired can be used to compare research papers and write smart reviews. In effect, the structured semantic representation of scientific knowledge in the KG makes it possible to automatically create literature comparisons. We are currently using these resources in our papers. One of these papers concerning "Food Composition Tables" is already published. The second one on "Food Information Engineering" was accepted at the AAAI conference.

5.6. Step 6: Verification and validation

The approach we present in this paper uses ORKG as an intelligent tool for assisting researchers in their work of organizing and comparing keyinsights extracted from existing literature. Thus, in step 4 and 5, we show how it can be used to create research contributions and compare scientific papers. To ensure that the templates, contributions, comparisons tables and smart reviews contain the necessary elements and that these elements are well structured and presented, they should be verified and validated. To this end, any researcher who has an account on the ORKG platform can edit any comparison, template, modify and save (for templates) or publish a new version (for comparisons).

6. Use cases

Knowledge acquired during the intervention phase of the Action research methodology presented in Section 4.2 were used to propose an approach using ORKG for knowledge acquisition from scientific papers (Section 5). This section constitutes the Post-Intervention of the Action research during which this methodology is used in real world settings to solve related problems. This approach was used to curate over 200 papers corresponding to the "ontology learning", "epidemiological surveillance systems design and implementation", "food information engineering", "Tabular data to Knowledge Graph Matching", "Question Answering", and "information extraction from scientific papers" research problems and, "Neuro-symbolic" domain⁵. From these research problems, we ingested over 800 contributions in the ORKG platform and we used these contributions to build over 100 comparisons tables. We used the template created during the curation of ORKG, and following steps 4 and 5 of the approach to create research contributions

⁵The overall work is freely available online at https://orkg.org/u/ebdd4856-0ac9-4a65-a077-470fe2ca6826 and https://orkg.org/u/aa79db4d-6762-4eb3-88fe-4db43405970c

of papers related to "ontology learning from text" and "ontology learning from videos". The "knowledge use" step consists of creating comparison tables of "ontology learning from videos" and "ontology learning from text" research problems. The overall links to all the resources presented in this Section are given as additional materials. The rest of this section presents how this approach was applied step by step to curate 21 papers related to epidemiological surveillance systems (Section 6.1), how this approach is currently used to curate papers in the domain of food information engineering (Section 6.2) and how we used it to curate the papers used to write the related work of this research (Section 6.3).

6.1. Epidemiological surveillance systems

Epidemiological surveillance systems enable the collection, analysis, and interpretation of data, together with the dissemination of these data to public health practitioners, clinicians, decision makers and the general population for preventing and controlling diseases [19, 20, 21]. It should support timely, efficient, flexible, scalable and interoperable data acquisition, analysis and dissemination. These informations are essential to the planning, implementation and evaluation of public health practices [19, 22]. To design and implement epidemiological surveillance systems, it can be important to have an overview of existing systems. Thus, this section presents how the approach presented in section 5 is used to acquire knowledge from papers related to epidemiological surveillance and build a comparison table.

6.1.1. Step 1: Knowledge elicitation

To furnish relevant information to stakeholders, epidemiological surveillance systems should be designed and implemented so as to always correspond to the requirements. Thus, the current work is about the acquisition of key-insights on epidemiological surveillance design and implementation with the goal to identify approaches, techniques and tools that are used for epidemiological surveillance and to see the limits of existing systems.

Given that epidemiological surveillance systems are primarily concerned with the collection, analysis, interpretation and dissemination of information to different stakeholders, we choose to classify the papers related to "Epidemiological surveillance systems design and implementation" research problem in the domain of "information science".

Once the research problem and the domain are identified, we move to the searching and the selection of papers that will be used. The famous research repository "Semantic Scholar" were used to search for relevant research papers: (1) "epidemiological surveillance system" search string were entered in the search bar of Semantic Scholar; (2) "Computer Science were chosen" as the field of study.

We found 44600 papers. We used the papers titles, short abstract provided by Semantic Scholar and paper abstract provided by the authors to select relevant papers. Given the large number of papers retrieved, we decided to consider only the first page of results provided by Semantic Scholar. Thereafter, we went through the papers on the first page one by one, selecting those that seemed to be relevant given the research problem. Citations-based analysis was also used to search for relevant papers. This consists of identifying all the papers that are cited and that have been cited. Fortunately these papers are extracted and automatically presented by Semantic Scholar. From the papers identified as relevant, a total of 21 were selected randomly and downloaded.

Once the papers were selected, we divided these papers into two groups: a group of 4 papers for building the template and the rest. Knowledge was acquired from these 4 papers by the identification of important key terms from each paper selected. Thus, each paper was read line by line, and we identified from each of them key-insights that may be of interest to researchers. After the elicitation phase, an exact and complete transcript of the key-insights extracted were made.

6.1.2. Step 2: Knowledge analysis and interpretation

Knowledge that was saved in the transcript was reviewed and analyzed in order to identify key pieces of knowledge and their relationships that represent scientific information carried by the research paper. A deep analysis of elements extracted were used to identify classes, properties and relations as described in Section 5. We were seeking the elements that are applicable to a considerable number of papers related to epidemiological surveillance systems.

6.1.3. Step 3: Template creation

The main classes and properties identified during Step 2 were used to build a template of papers related to "epidemiological surveillance systems design and implementation⁶" research problem. This template is available online and can be improved by other researchers.

⁶https://www.orkg.org/orkg/template/R150089

6.1.4. Step 4: Knowledge representation

During the knowledge representation step, the template built throughout the previous step was used to annotate the 4 papers used to build the template. Thereafter, knowledge were acquired from the rest of papers and ingested in ORKG using the template. In total, 21 papers were ingested in ORKG.

6.1.5. Step 5: Knowledge use

The contributions created using the template were used to build a comparison table⁷. The latter compares papers related to the "epidemiological surveillance system design and implementation" research problem.

6.1.6. Step 6: Knowledge verification and validation

The discussion with a collaborator who is an epidemiologist allowed us to validate the template. They found that the template, the comparison table and the contributions constitute the elements that are helpful when putting in place an epidemiological surveillance system.

6.2. Food Information Engineering

Food information engineering involves the acquisition, the processing and the diffusion of up-to-date food information to different stakeholders. These informations are compiled from several data sources and used for a variety of purposes such as food recommendation, recipe substitution, food image recognition, etc. Many authors have proposed methodologies, methods and tools for the acquisition, the processing of food information, its storage, diffusion, etc. However, these contributions are scattered in many scientific papers on the Internet and are difficult to exploit. The second use case we chose consists of documenting the "food information engineering" research problem so as to provide to fellow researchers with methodologies, methods, tools, use cases, etc. It consists of documenting the following research question: "how food information is collected, processed, diffused and used?" To reply to this research question, several researches on the acquisition of food knowledge, its storage, querying and diffusion to different stakeholders are done worldwide. Our objective during this work is to document these solutions so as to provide to the research community with a body of knowledge that will help fellow researchers to reduce their research curve.

⁷https://www.orkg.org/orkg/comparison/R146851/

6.2.1. Step 1: Knowledge elicitation

Our goal during the Knowledge elicitation step was to identify several papers that can allow us to document the research question: "how food information is collected, processed, diffused and used?" We were having prior knowledge on the organization of food information using Food Composition Tables, Food Ontologies and Food Knowledge Graphs. Thus, we position food information engineering research problem in the Semantic Web research domain. Once the research problem to work on and the research domain was determined, we move to the searching of relevant papers. Our goal being to build comparison tables for the following research problems:

- Food Composition Tables construction and description (5 papers processed, 4 comparisons tables built),
- Food Ontology construction, description and integration (27 papers processed, 4 comparison tables built and one smart review wrote),
- Food knowledge graph construction, description and integration (11 papers processed, 4 comparisons tables built and one smart review wrote),
- etc.

We used Google search to search for relevant papers from these subjects, using as keywords the title of the research problems. Only the first page (containing 10 research results) of the Google search platform was considered. In the case of "Food Ontologies" and "Food Knowledge Graphs", we used the most recent review published by Weiqing et al. [23] to identify the research papers related to "Food Ontologies" and "Food Knowledge Graphs". Once retrieved, we choose some of these papers to identify elements that are comparable.

6.2.2. Step 2: Knowledge analysis and interpretation

As we did with epidemiological surveillance systems, knowledge that was identified from the papers downloaded and saved in transcript was reviewed and analyzed in order to identify classes, properties and relations. We used the comparison of "Food Ontologies" and "Food Knowledge Graphs" provided by Weiqing et al. [23] to find additional properties. These tables were imported in the ORKG system⁸⁹. In this particular case, we started by

⁸https://orkg.org/comparison/R221127

⁹https://orkg.org/comparison/R217515/

getting and using all properties used to compare papers in the review paper.

During the analysis of papers, we found that many authors were providing Question Answering Systems over Food KG. Thus, we decided to document this research problem¹⁰¹¹.

6.2.3. Step 3: Templates creation

Once the papers were selected, they were used to build new templates (8 templates) and update existing templates (two templates were updated). For instance, the template of "Ontology description" created during the Intervention phase (Section 4.2) was updated with new properties, another template created by Natalia Chichkova, an ORKG user for the description of KG was also updated. The following examples of templates were created by zero and are currently used: food composition tables, Question Answering systems and Question Answering benchmark.

6.2.4. Step 4: Knowledge representation

During the knowledge representation step, the template built throughout the previous step was used to annotate all the papers downloaded by considering different research problems. Currently, more than 120 papers related to the domain of "Food Information Engineering" are ingested in the ORKG platform. It should be noted that this is an ongoing work and we want at its end to provide to the research community with a systematic literature review of "food information engineering" research problems.

6.2.5. Step 5: Knowledge use

The contributions created using the template were used to build over 26 comparison tables. The comparison table of "food composition table" research papers allowed us to realize that "food composition tables" change over time and unfortunately, the database did not change. On the other hand, the supports used to distribute these data are sparse on the Internet in different formats. We also realize that up-to-date data can be found in scientific papers. Thus, we build a large scale and up-to-date food composition tables that is currently annotated using Wikidata.

6.2.6. Step 6: Knowledge verification and validation

Knowledge validation consists of presenting this work in challenges and conferences. Our work on "Food Composition Table" was accepted in SemTab

¹⁰https://orkg.org/comparison/R239314

¹¹https://orkg.org/comparison/R269002

challenge 12 organized by International Semantic Web Conference 2022^{13} . The overall work on food information engineering was accepted at "New Faculty Highlights" AAAI- 23^{14} conference program. We are currently adding more papers in order to maintain a state of the art of papers in the domain of "food information engineering".

6.3. Knowledge extraction from scientific papers

The process of literature review starts from the searching of scientific papers from the huge amount of existing ones to the analysis of the paper content and the extraction of key-insights from them. Given the large amount of scientific papers in all domains, this process is laborious, time consuming and cumbersome. To reduce the burden of work, knowledge extraction from scientific papers is of great interest to researchers. During the last years, this research problem has interested many researchers and methodologies, methods and tools have been proposed. Our goal during the third use case was to identify the different types of knowledge that are extracted from scientific papers and to document datasets, methodologies, models and tools used for extracting these knowledge.

6.3.1. Step 1: Knowledge elicitation

Given that the research problem we are documenting is "knowledge extraction", we classified this research problem in the Semantic Web domain. As we did with the two previous use cases, our goal during this step was to identify several papers that can cover the research question we want to document. By using the search keyword: "knowledge extraction from scientific paper" on Google Search engine, we found a great survey [4]. This is a 60 pages survey of datasets, methodologies, methods and tools that are used to extract different types of knowledge from scientific papers. It is organized in two main sections: (1) Metadata extraction, (2) Key-insights extraction. Each section describes the different types of knowledge that are extracted, the methods that are used to extract each type of knowledge and the evaluation of each method. We found this survey interesting for knowledge elicitation.

The survey paper was read line by line in order to identify elements that are comparable. The comparisons tables provided by the authors were great resources for the identification of key-insights. Thus, we combine the

¹²https://sem-tab-challenge.github.io/2022/

¹³https://iswc2022.semanticweb.org/

¹⁴https://aaai.org/Conferences/AAAI-23/new-faculty-highlights-cfp/

knowledge extracted from these tables with the knowledge extracted from the full body text to obtain a set of key-insights candidates.

6.3.2. Step 2: Knowledge analysis and interpretation

The key-insights identified from the tables and the text were analyzed one by one in order to select the ones that can be considered as relevant. The duplicates were also identified and deleted.

6.3.3. Step 3: Templates creation

Knowledge identified during the previous step were converted into properties, classes and relations. Thereafter, these classes, properties and relations were used to create templates. We found it necessary the create the following templates:

- Template for metadata dataset¹⁵: this template is used to describe the content of each metadata dataset.
- Templates for Key-Insight¹⁶¹⁷: two types of datasets describing keyinsights were found: sentence-level key-insight and phrase-level keyinsights. These templates are used to describe these datasets.
- Template of metadata system¹⁸: this is used to describe the different systems that are used for extracting the metadata from the scientific article.
- Template of key-insight system¹⁹: this is used to describe the different systems that are used for extracting key-insights from scientific papers.

In addition to these templates, we reused a template²⁰ that we created during the work on "food information engineering" for evaluating each extraction system. We also used a template²¹, created by Jennifer D'Souza for the description of existing tools that are proposed for knowledge extraction from scientific papers.

¹⁵https://orkg.org/template/R277000

¹⁶https://orkg.org/template/R279223

¹⁷ https://orkg.org/template/R280533

¹⁸ https://orkg.org/template/R280212

¹⁹ https://orkg.org/template/R280523

²⁰ https://orkg.org/template/R259041

²¹ https://orkg.org/template/R166722

6.3.4. Step 4: Knowledge representation

During the knowledge representation step, the template built throughout the previous step was used to annotate papers related to "information extraction from scientific papers".

6.3.5. Step 5: Knowledge use

Currently, more than 50 papers related to "information extraction from scientific papers" are being ingested in ORKG. These papers are used to document the "information extraction from scientific papers" research problem. From these papers, more than 50 research contributions were extracted and used to build 11 comparison tables. These resources were used to write the related work of this research (see Section 7).

6.3.6. Step 6: Knowledge verification and validation

The templates and the contributions provided in this research will be evaluated by the reviewers of this paper. On the other hand, these resources can be evaluated, validated and improved by any researcher working on knowledge extraction from scientific papers.

7. Related work

As presented in the previous sections, scientific knowledge can be grouped into two categories: metadata and key-insights [4, 5].

During the last years, many researchers have contributed in the domain of metadata extraction from research papers. Zara et al [4] and Abdul et al. [5] present a great state-of-the-art on this subject. These works show that manual processing is generally used for scientific papers annotations in order to build datasets. Thereafter, these datasets are used to train models that will further be used for metadata extraction. The models that are used for metadata extraction are rule-based, machine learning-based and Natural Language Processing-based. Rule-based models use text features and layouts to define instructions that specify how to extract desired information from scientific papers. On the other hand, methods such as Hidden Markov Models (HMM), Conditional Random Fields (CRF), Support Vector Machines (SVM), Neural Networks are also proposed for metadata extraction from scientific papers. The approaches proposed for metadata extraction are very powerful. The evaluation of the most powerful ones show the performances reaching 95% of F-measure.

Key-insights acquisition consists of reading the scientific paper, identifying relevant knowledge and organizing them or building models for their automatic extraction. In the rest of this section, we present the different types of key-insights in Section 7.1, existing key-insights datasets in Section 7.2, methods for key-insights extraction in Section 7.3, and tools for key-insights extraction in Section 7.4.

7.1. Key-insights

Key-insights are presented in scientific papers in the form of text, figures and tables. The semi-structured organization of knowledge in tabular data allows us to easily extract key-insights from tables stored in scientific papers. For instance, Food Composition Tables can be extracted in scientific papers for accessing food that people are eating and their nutritive values [24]. However, key-insights hidden in text are more difficult to identify and extract because it is difficult to guess the valuable information enclosed within a research paper text that can be beneficial for each researcher. Zara et al. [4] classified key-insights hidden in the paper text into sentence-level key-insights, phrase-level key-insights, and relation [4]:

- Sentence-level key-insights: these are predefined knowledge, in the form of keywords and key-phrases and hidden in the text of an article. For instance, "method", "problem", "objective", "result", etc. are included in almost all scientific papers.
- Phrase-level key-insights: These are phrases carrying potential information that are useful to researchers. For instance, "tool or library", "measures and measurements", "language resource product", "location", etc.
- Relation: relation can express application of a technique to solve a problem, results generated against various evaluation measures, etc. Phrase-level key-insights can be extended to extract relations because in many cases, relations are expressed between entities.

Key-insights acquisition from scientific papers can be done manually or automatically. We presented in Sections 3, 4, 5 and 6 how ORKG can be used as a computer-assistant tool for semi-automatic acquisition of knowledge from scientific papers. To build models for automatic acquisition (or extraction) of key-insights from scientific papers, there is a need for annotated datasets. In the next section, we present related work on key-insights datasets.

7.2. Datasets

Based on the different types of key-insight that can be extracted from scientific papers, the datasets for extracting these knowledge can be classified as Sentence-level key-insights and Phrase-level key-insight datasets.

- Sentence-level key-insights datasets: These datasets contain scientific articles in which sentences are classified based on insights they carry. We gathered the different properties that can be used to compare sentence-level key-insights and we built an ORKG template. Thereafter, this template was used to compare Sentence-level key-insights datasets published in scientific literature.
- Phrase-level key-insight datasets: these datasets contain scientific papers in which phrases are annotated with entities corresponding to potential key-insights they may carry. The datasets for phrase-level key-insight extraction are difficult to build and scarce. As we did with sentence-level key-insights, we built an ORKG template of phrase-level key-insights and we used this template to compare phrase-level key-insights datasets.

The comparison of phrase-level and sentence-level key-insights shows that the majority of existing datasets belong to the domain of medical science. On the other hand, these datasets are mainly based on the extraction of knowledge from the abstract only [4].

7.3. Acquisition methods

Acquiring knowledge from scientific papers can be manual or automatic. Automatic knowledge acquisition relies on rules, Machine Learning, Deep Learning and Natural Language Processing techniques for automatic identification and extraction of key-insights. Based on the datasets presented in Section 7.2, Zara et al. [4] classified these methods as sentence-level key-insights and phrase-level key-insights methods.

- Sentence-level key-insights extraction: methods for Sentence-level key-insights extraction are focused on the classification of sentences in predefined categories based on insights they carry.
- Phrase-level key-insights extraction: methods for Phrase-level key-insights extraction are focused on the extraction of phrases carrying potential information.

To extract sentence-level and phrase-level key-insights from scientific papers, rules, ML, DL and NLP techniques have been proposed. The main techniques proposed are Bayesian classifier, Conditional Random Field, Support Vector Machine, Hidden Markov Models. To compare research work on this subject, we built a template and we used this template to compare several methods for sentence-level and phrase-level key-insights extraction. These methods are not as powerful as metadata extraction methods. Very little works show methods that the F-measure reaches 85% for the extraction of each key-insights.

7.4. Tools for knowledge acquisition

Key-insights acquired from scientific papers are generally grouped into research contributions to make them comparable with other resources. To this end, hand-written notes can be used to organize and build comparison tables and figures. Tools used for knowledge acquisition from scientific papers aim to facilitate this work and make it less laborious, time consuming and cumbersome. They can be classified as computer-assisted tools, tools for automatic extraction of key-insights, digital research repositories and social tagging and bookmarking platforms.

7.4.1. Computer-assisted tools

Computer-assisted tools aim to help researchers to organize key-insights extracted from scientific papers. Spreadsheets software such as Microsoft Excel, Libreofficel Cal or Google spreadsheets are generally used to organize, store and compare research contributions from several research papers. The main advantages of these software is that the data can be stored and reused whenever needed. It is also easier to build graphics with the data. However, these data are not harmonized, isolated in researcher computers or storage and difficult to merge with other research data. Thus, two researchers will make the same effort to extract the same knowledge from a set of scientific papers. These efforts can be saved if the knowledge is organized in a computer-assisted software such as ORKG. In a recent work, Allard et al. [13] present a workflow designed to compare research contributions in ORKG. This paper shows the process to add a paper and the key-insights of this paper in ORKG. However, it did not provide a complete methodology from the knowledge elicitation phase (using template to create a conceptual model of the domain) to the knowledge use phase.

7.4.2. Digital research repositories

Digital research repositories aim at providing researchers with basic filters to ease the search of scientific papers while querying through millions of research papers. To this end, metadata informations are used to provide various searching facilities [4]. On the other hand, key-insights are used to augment keywords and provide short abstracts (e.g., Semantic Scholar) to guide researchers to identify relevant papers to his research problem.

7.4.3. Social tagging and bookmarking platforms

Social tagging and bookmarking platforms (e.g. CiteULike, Bibsonomy, Delicious) are online services for serving scientific communities [5]. The users of these tools can annotate the research articles, bookmark the preferences, etc. This allows them to possess their references or a web page with their own defined tags or keywords. But this does not allow researchers to compare research contributions identified from several research papers.

Even if the knowledge of some Digital research repositories and Social tagging and bookmarking platforms are organized in knowledge graphs (e.g., Springer Nature SciGraph²², Microsoft Academic [25]), these tools does not permit to researchers to structure key-insights hidden so to help other researchers to update with more papers and insights.

8. Summary and conclusion

Acquiring knowledge from scientific papers from scratch is costly in time and resources. Thus, we propose in this paper an approach using Open Research Knowledge Graph as a computer-assistant tool for knowledge acquisition from scientific papers. It consists of five steps:

- Knowledge elicitation consists of determining the domain and the research problem to document. Using these information, to search for relevant scientific papers and extract elements that one wants to compare.
- Knowledge analysis and interpretation consist of analyzing the pertinence of the elements extracted during knowledge elicitation and the deletion of duplicates.
- Template creation consists of using the elements obtained after the knowledge analysis and interpretation to build a template that will be

²²https://www.springernature.com/gp/researchers/scigraph

used further to organize key-insights extracted and research contributions.

- Knowledge representation consists of using existing templates to structure knowledge extracted in a knowledge graph.
- Knowledge use consists of comparing research contributions in comparison tables, and using them to write reviews of the domain.
- Verification and validation consists of the validation of the templates, the contributions, the comparisons of research contributions and the reviews by fellow researchers.

This approach is currently used to document the "ontology learning", "epidemiological surveillance systems design and implementation", "food information engineering", "Tabular data to Knowledge Graph Matching", "Question Answering", and "information extraction from scientific papers" research problems and the "Neuro-symbolic AI" domain. Thus, more than 200 papers are ingested in ORKG. From these papers, more than 800 contributions are documented and these contributions are used to build over 100 comparison tables. At the end of this work, we found that ORKG is a valuable tool that can reduce the working curve of state-of-the-art research.

Acknowledgement

We are grateful to the Open Research Knowledge Graph team for their following during the curation of ORKG. Our great thanks also goes also to all the curators. Their remarks and questions were very helpful in this work.

References

- [1] K. Jayaram, K. Sangeeta, A review: Information extraction techniques from research papers, in: 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017, pp. 56–59. doi:10.1109/ICIMIA.2017.7975532.
- [2] P. K. Shah, C. Perez-Iratxeta, P. Bork, M. A. Andrade, Information extraction from full text scientific articles: Where are the keywords?, BMC Bioinform. 4 (2003) 20. URL: https://doi.org/10.1186/1471-2105-4-20. doi:10.1186/1471-2105-4-20.

- [3] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. Eddine Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, BIBLIOTHEK – Forschung und Praxis (2020). doi:http://dx.doi.org/10.18452/22049.
- Z. Nasar, S. W. Jaffry, M. K. Malik, Information extraction from scientific articles: a survey, Scientometrics 117 (2018) 1931–1990.
 URL: https://doi.org/10.1007/s11192-018-2921-5. doi:10.1007/s11192-018-2921-5.
- [5] A. Shahid, M. T. Afzal, M. Abdar, M. E. Basiri, X. Zhou, N. Y. Yen, J. Chang, Insights into relevant knowledge extraction techniques: a comprehensive review, J. Supercomput. 76 (2020) 1695–1733. URL: https://doi.org/10.1007/s11227-019-03009-y. doi:10.1007/s11227-019-03009-y.
- [6] K. Adnan, R. Akbar, Limitations of information extraction methods and techniques for heterogeneous unstructured big data, International Journal of Engineering Business Management 11 (2019) 1847979019890771. doi:10.1177/1847979019890771.
- [7] E. Acheson, R. S. Purves, Extracting and modeling geographic information from scientific articles, PLOS ONE 16 (2021) 1–19. URL: https://doi.org/10.1371/journal.pone.0244918. doi:10.1371/journal.pone.0244918.
- [8] D. G. Caroline Hayes, Michael Fu, Some Guidelines for Knowledge Acquisition Strategies, Technical Report, AAAI Technical Report SS-96-02, 1996. URL: https://www.aaai.org/Papers/Symposia/Spring/1996/SS-96-02/SS96-02-008.pdf.
- [9] D. Gasevic, D. Djuric, V. Devedzic, Model Driven Engineering and Ontology Development, 2nd ed., Springer Publishing Company, Incorporated, 2009.
- [10] J. W. Karl, J. E. Herrick, R. S. Unnasch, J. K. Gillan, E. C. Ellis, W. G. Lutters, L. J. Martin, Discovering Ecologically Relevant Knowledge from Published Studies through Geosemantic Searching, BioScience 63 (2013) 674–682. doi:10.1525/bio.2013.63.8.10.
- [11] J. D. Margulies, N. R. Magliocca, M. D. Schmill, E. C. Ellis, Ambiguous geographies: Connecting case study knowledge with global change

- science, Annals of the American Association of Geographers 106 (2016) 572–596. doi:10.1080/24694452.2016.1142857.
- [12] M. N. Kamel, Knowledge acquisition, in: B. W. Wah (Ed.), Wiley Encyclopedia of Computer Science and Engineering, John Wiley & Sons, Inc., 2008. doi:10.1002/9780470050118.ecse205.
- [13] A. Oelen, M. Y. Jaradeh, K. E. Farfar, M. Stocker, S. Auer, Comparing research contributions in a scholarly knowledge graph, in: D. Garijo, M. Markovic, P. Groth, I. Santana-Pérez, K. Belhajjame (Eds.), Proceedings of the Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), Marina del Rey, California, November 19th, 2019, volume 2526 of CEUR Workshop Proceedings, 2019, pp. 21–26.
- [14] R. et al., Empirical standards for software engineering research, 2020. URL: https://arxiv.org/abs/2010.03525. doi:10.48550/ ARXIV.2010.03525.
- [15] F. J. Azanzi, G. Camara, M. Tchuente, Extracting ontological knowledge from java source code using hidden markov models, Open Computer Science 9 (2019) 181–199. URL: https://doi.org/10.1515/comp-2019-0013. doi:10.1515/comp-2019-0013.
- [16] A. Konys, Knowledge systematization for ontology learning methods, in: Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference KES-2018, Belgrade, Serbia, 3-5 September 2018., 2018, pp. 2194–2207. URL: https://doi.org/10.1016/j.procs.2018.07.229.doi:10.1016/j.procs.2018.07.229.
- [17] M. Shamsfard, A. Abdollahzadeh Barforoush, The state of the art in ontology learning: A framework for comparison, Knowl. Eng. Rev. 18 (2003) 293–316.
- [18] L. Zhou, Ontology learning: state of the art and open issues, Information Technology and Management 8 (2007) 241–252. URL: https://doi.org/10.1007/s10799-007-0019-5. doi:10.1007/s10799-007-0019-5.
- [19] B. C K Choi, The past, present, and future of public health surveillance, Scientifica 2012 (2012) 875253.

- [20] C. L. Richards, M. F. Iademarco, D. Atkinson, R. W. Pinner, P. Yoon, W. R. M. Kenzie, B. Lee, J. R. Qualters, T. R. Frieden, Advances in public health surveillance and information dissemination at the centers for disease control and prevention, Public Health Reports 132 (2017) 403–410.
- [21] A. Jiomekong, G. Camara, Model-driven architecture based software development for epidemiological surveillance systems, Studies in health technology and informatics 264 (2019) 531—535. URL: https://doi.org/10.3233/SHTI190279. doi:10.3233/shti190279.
- [22] R. R. Frerichs, Epidemiologic surveillance in developing countries, Annual Review Public Health 12 (1991) 257.
- [23] W. Min, C. Liu, L. Xu, S. Jiang, Applications of knowledge graphs for food science and industry, Patterns 3 (2022) 100484. doi:https://doi.org/10.1016/j.patter.2022.100484.
- [24] A. Jiomekong, C. Etoga, B. Foko, V. Tsague, M. Folefac, S. Kana, M. M. Sow, G. Camara, A large scale corpus of food composition tables, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), CEUR-WS. org (2022).
- [25] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft Academic Graph: When experts are not enough, Quantitative Science Studies 1 (2020) 396–413. doi:10.1162/qss_a_00021.