knowledge

MDPI

Article

Agriculture Named Entity Recognition—Towards FAIR, Reusable Scholarly Contributions in Agriculture

Jennifer D'Souza

TIB Leibniz Information Centre for Science and Technology, 30167 Hannover, Germany; jennifer.dsouza@tib.eu

Abstract: We introduce the Open Research Knowledge Graph Agriculture Named Entity Recognition (the ORKG Agri-NER) corpus and service for contribution-centric scientific entity extraction and classification. The ORKG Agri-NER corpus is a seminal benchmark for the evaluation of contribution-centric scientific entity extraction and classification in the agricultural domain. It comprises titles of scholarly papers that are available as Open Access articles on a major publishing platform. We describe the creation of this corpus and highlight the obtained findings in terms of the following features: (1) a generic conceptual formalism focused on capturing scientific entities in agriculture that reflect the direct contribution of a work; (2) a performance benchmark for named entity recognition of scientific entities in the agricultural domain by empirically evaluating various state-of-the-art sequence labeling neural architectures and transformer models; and (3) a delineated 3-step automatic entity resolution procedure for the resolution of the scientific entities to an authoritative ontology, specifically AGROVOC that is released in the Linked Open Vocabularies cloud. With this work we aim to provide a strong foundation for future work on the automatic discovery of scientific entities in the scholarly literature of the agricultural domain.

Keywords: information extraction; named entity recognition; natural language processing; dataset; sequence labeling; scholarly knowledge graphs; open research knowledge graph

Citation: D'Souza, J. Agriculture Named Entity Recognition—Towards FAIR, Reusable Scholarly Contributions in Agriculture. *Knowledge* **2024**, *4*, 1–26. https://doi.org/10.3390/knowledge4010001

Academic Editor: Constantin Bratianu

Received: 16 May 2023
Revised: 11 January 2024
Accepted: 15 January 2024
Published: 19 January 2024

1. Introduction

Scientific innovations drive progress in companies, industries and the economy. Currently, the scholarly publication cycles are at an alarming rate of 2.5 million articles per year [1]. Thus, the traditional documents ranked lists offered by scholarly search engines no longer support efficient research and development (R&D). While they pinpoint individual papers of interest from a mass of documents, they do not offer researchers a sense of an overview of the field. Researchers seem to drown in the deluge of publications as a consequence of the tediously long information assimilation cycle to manually scan salient aspects of research contributions within information buried in static text. Thus, enabling machine-actionability of scholarly knowledge is warranted now more than ever. In this vein, the method of scholarly knowledge strategic reading powered by Natural Language Processing (NLP) is being advocated for research, business, government, and non-governmental organization (NGO) stakeholders [2]. Most current strategic reading relies on human recognition of scientific terms from text, perhaps assisted with string searching and mental calculation of ontological relationships, combined with burdensome tactics of bookmarking, note-taking, and window arrangement. To this end, recently, an increasing number of research efforts are geared toward putting in place next-generation Findable, Accessible, Interoperable, and Reusable (FAIR) [3] scholarly knowledge representation models as Knowledge Graphs (KGs) [4,5]. They advocate advanced semantic machine-interpretability of publications via KGs to enable more intelligent automated processing (e.g., smart information access). This development started in advanced scholarly digital libraries (DL) such as the Open Research Knowledge Graph (ORKG, https://orkg.org/,

accessed on 14 January 2024) [5], that crowdsources templated research contributions resulting in tabulated surveys of comparable contributions (cf. Figure 1), thus demonstrates strategic reading in practice.

Properties	6th Open Challenge on Question Answering over Linked Data (QALD-6) *Contribution 4 - 2016*	6th Open Challenge on Question Answering over Linked Data (QALD-6) *Contribution 3 - 2016*	6th Open Challenge on Question Answering over Linked Data (QALD-6) *Contribution 2 - 2016*
has research problem	Question answering systems evaluation	Question answering systems evaluation	Question answering systems evaluation
On	QALD-6	QALD-6	QALD-6
dataset	DBpedia 2015	DBpedia 2015	DBpedia 2015
	DBpedia 2015 with abstracts	DBpedia 2015 with abstracts	DBpedia 2015 with abstracts
	LinkedSpending	LinkedSpending	LinkedSpending
evaluation	SemGraphQA	UTQA	UTQA
Language	Farsi	English	Spanish
F-measure	0.37	0.65	0.68
Precision	0.70	0.70	0.76
Question amount	100	100	100
Recall	0.25	0.61	0.62

Figure 1. Demonstration of strategic reading of machine-actionable representations of 3 scholarly contributions on the problem of Question Answering in the Open Research Knowledge Graph (ORKG), which generates aggregated comparative views from the graph-based semantic research contribution descriptions.

To represent scholarly publications as KGs, from an Information Extraction (IE) perspective, named entity recognition (NER) over scholarly publications becomes a vital task since entities are at the core of KGs. As an IE task, NER over scholarly documents is a long-standing task in the NLP community–the Computer Science domain itself has been addressed over a wide body of works with various knowledge capture objectives [6–19]. However, this well-established research area [20–23], thus far, has not seen any practical applications in the Agricultural scholarly publications domain.

In the domain of agriculture, the gradual sophistication of food production and agricultural methods led to an increasing demand for data exchange, processing and information retrieval. Thus the recording of knowledge as information islands via manual notetaking had to evolve to the recording of relational knowledge in databases via protocols. These protocols facilitated standardized recording and exchange of knowledge between different databases via purposefully invented data dictionaries and coding systems that assigned simple alphanumeric codes to products, varieties, breeds or crops. E.g., the ISOBUS [24]/ISO11783 [25] data dictionary or the European and Mediterranean Plant Protection Organization (EPPO) codes of crops used for plant protection applications [26]. Today, however, we are faced with not only sophisticated agricultural practices but also voluminous masses of agricultural research findings published worldwide. Hence the call for the adoption of next-generation semantic web publishing model [27] of machine-actionable, structured scholarly contributions content via the ORKG platform. Within this model, a large-scale agricultural KG would be predicated on standardized templated subgraph patterns for recording interoperable structured scholarly contributions in agriculture. The custom-templated subgraphs ensure the standardized recording of *comparable* research

contributions in an overarching interoperable graph of highly varied underlying research domains. The research domains can include appraisals of agricultural products, e.g., A chemotaxonomic reappraisal of the Section Ciconium Pelargonium (Geraniaceae) [28], or the restoration and management of plant systems, e.g., mangrove systems. Table 1 lists 15 sub research domains of contemporary research in agriculture. An information modeling objective ensures capturing contributions under a uniform set of salient properties within a single domain, while allowing for the definition of varied sets of salient properties across domains. This enables machine-assisted strategic reading within the semantic web publishing model directly addressing the information ingestion problem over massive volumes of findings for the researchers by smart machine assistance. E.g., as structured contribution comparisons computed over the set of salient contribution properties in one domain as depicted in Figure 1.

Table 1. A listing of 15 different research domains in the table columns that were observed in a corpus of 5500 scholarly publication titles in Agriculture.

Agriculture Research Domains		
fertilizers	different natures of agricultural production communities such as the winter-rainfall desert community	restoration and management of plant systems
microalgae refineries	first reports on plant findings	investigation of biochemical activities in plant species
climatic factors such as fires affecting food production	appraisals of agricultural products	in vitro cultivation of plant species
land degradation and cultivation research	competitive growth advantages of paired cultivation	characterizing seeds or plant species
antibacterial and chemical byproducts from plants	creation of taxonomic lists of crops	importing new plant species across regions

The road to discovering contribution templates for research domains should be based on a set of generic entity types being applicable across all domains that can be further specialized and instantiated as domain-specific, full-fledged templates. In other words, prior to obtaining research-domain-specific contribution template patterns, there needs to be put in place a standardized set of generic entity types that can foster the further development of the problem-specific contribution templates constituted by additional semantic properties. As such the Agriculture Named Entity Recognition service of the ORKG (the ORKG Agri-NER service), addressed seminally in this work, proposes a set of seven generic entity types that encapsulate the contribution of a work extracted from paper titles. The seven *contribution-centric* entity types are: RESEARCH PROBLEM, RESOURCE, PROCESS, LOCATION, METHOD, SOLUTION, and TECHNOLOGY. Building on this idea, this study makes two novel key contributions: (1) we propose for the first time an NER service specifically tailored for the agricultural domain; and (2) predicated on seven contribution-centric entities derived from paper titles and inspired from the top-level concepts of the AGROVOC ontology (https://agrovoc.fao.org, accessed on 14 January 2024) of the Food and Agriculture Organization of the United Nations (FAO, https://www.fao.org/home/en/, accessed on 14 January 2024), we lay the groundwork for the discovery of domain-specific contribution templates for the further specification of the generic entity types.

The ORKG Agri-NER service is an IE system of seven entity types such as research problems, resources, location of study, etc., which since extracted from paper titles implicitly encapsulate the contributions of scholarly articles. Conceptually, the shared understanding around paper titles is that they are succinct summarizations of the contribution of a work [18]. Thus when looking to formulate a contribution-centric entity extraction objective, the first place to seek out this information is from paper titles. Specifically, ORKG Agri-NER provides a conceptual ecosphere of seven entity types to begin to generically structure and

compare the contributions of scholarly articles in the domain of Agriculture as illustrated in Figure 2. A striking feature of the proposed work is that it supports retrieving, exploring and comparing research findings based on explicitly named entities of the knowledge contained in agricultural scientific publications. If applied widely, ORKG Agri-NER can have a significant impact on scholarly communication in the agricultural domain. It specifically addresses researchers who want to compare their research with related works, get an overview of works in a certain field, or search for research contributions addressing a particular problem or having certain characteristics. Figure 3 gives a high-level overview of the proposed semantic model by showing the seven core entity types in Agri-NER. The ORKG Agri-NER service then is the first step in a long-term research agenda to create a paradigm shift from document-based to structured knowledge-based scholarly communication for the agricultural domain. Other than the discovery of contribution-centric template patterns in the ORKG, the machine-readable description of research knowledge in the seven entity types could support other services for analyzing scientific literature in the agricultural domain such as forecasting agricultural research dynamics, identifying key insights, informing funding decisions, and confirming claims in news on contemporary agricultural research. To facilitate further research, we contribute two resources to the community: (1) The ORKG Agri-NER human-annotated gold-standard corpus which can be downloaded at https://github.com/jd-coderepos/contributions-ner-agri (accessed on 14 January 2024) under the CC BY-SA 4.0 license; and (2) The ORKG Agri-NER tool whose source code can be accessed at https://gitlab.com/TIBHannover/orkg/nlp/experiments/orkg-agriculture-ner (accessed on 14 January 2024) under the MIT license, and which furthermore are available as services to the community in two ways–the python package version of the service can be accessed at https://orkg-nlp-pypi.readthedocs.io/en/latest/services/services.html (accessed on 14 January 2024); also, it is possible to directly interact with the REST API for the Agri-NER service directly via the interaction documentation page at https://orkg.org/nlp/api/docs#/annotation/annotates_agri_paper_annotation_agriner_post (accessed on 14 January 2024). The remainder of the paper explains both the creation of the dataset resource and tool in detail.

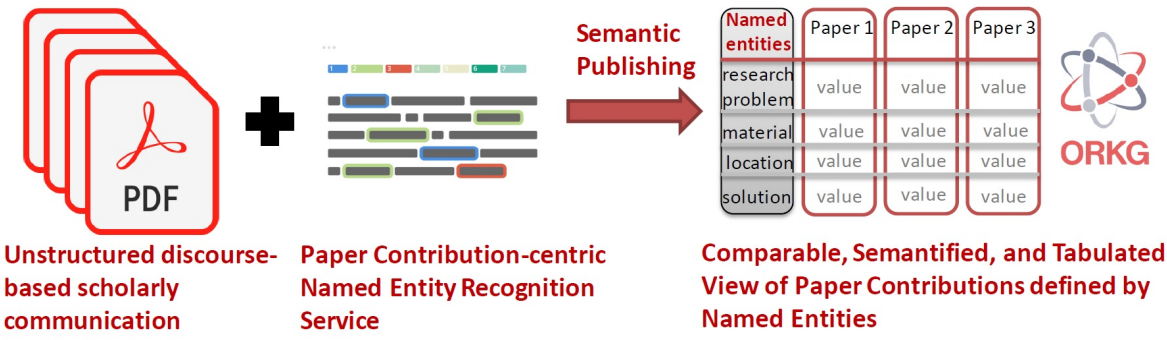

Figure 2. Transition from document-based to contribution-centric named entity recognition service-powered knowledge-based scholarly communication.

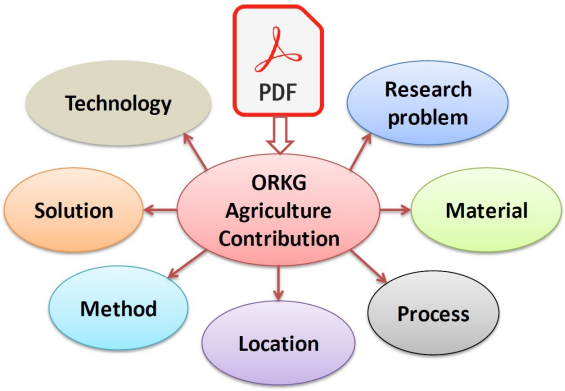

Figure 3. The seven core concepts proposed to capture contributions from scholarly articles in the Agriculture domain.

2. Background

> "Semantic Web ... does not require complex artificial intelligence to interpret human ideas, but 'relies solely on the machine's ability to solve well-defined problems by performing well-defined operations on well-defined data' ". [27,29]

The FAIRification guidelines [3] for scholarly knowledge publishing inadvertently advocates for adopting semantic models for machine-actionable knowledge capture of certain aspects of the article *content* such that they are findable, actionable, interoperable, and reusable. Ontological or entity-centric conceptual schemas are an elegant demonstration of what "going FAIR" (https://www.go-fair.org/fair-principles/, accessed on 14 January 2024) means in practice across the broad spectrum of the researchers landscape as long as they are involved in the publication of their work. These schemas, by going beyond 'data' in the conventional sense, and instead applying to algorithms, tools, and workflows that lead to the data which are traditionally captured in discourse text, bring the recording of these aspects of scholarly knowledge in the FAIR landscape. Thereby, transparency, reproducibility, and reusability of scholarly analytical pipelines are fostered. Broadly, the research paradigms around the generation of FAIR data can be classified into two broad types: (1) ontological models that can directly produce FAIR-compliant data when instantiated; and (2) informal conceptual annotation models which are not characteristically FAIR-compliant but which work on data instances that support the discovery of ontologies in a bottom-up manner. These models equip experts with a tool for semantifying their scholarly publications ranging from strictly-ontologized methodologies [30,31] to less-strict, flexible conceptual description schemes [7,17], wherein the latter aim toward the bottom-up, data-driven discovery of an ontology.

The remainder of this section is organized per these two broad paradigms.

2.1. Ontological Structuring of Scholarly Publications

Early works can be traced to the Dublin Core Metadata Terms (DCTerms) [32] ontology (http://purl.org/dc/terms/, accessed on 14 January 2024). The original "Dublin Core" was the result of a March 1995 workshop in Dublin, Ohio, which sought to define a generic metadata record that was generic enough to describe a wide range of electronic objects [33]. Subsequent ontologies specifically modeled scholarly articles but inherited DCTerms in an upper-level ontology space.

Some ontologies focused on modeling the scholarly document structure and rhetorics. In this vein, the Document Components Ontology (DoCO) [34] is an ontology for describing both structural and rhetorical document components in RDF. For structural annotations, DoCO imports the Document Structural Patterns Ontology (https://sparontologies.github.io/po/current/po.html, accessed on 14 January 2024) with classes such as Sentence, Paragraph, Footnote, Table, Figure, CaptionedBox, FigureBox, List, BibliographicReferenceList etc. The pattern ontology defines formally patterns for segmenting a document into atomic

components, in order to be manipulated independently and reflowed in different contexts. For the rhetorical annotations, DoCO imports the Discourse Elements Ontology (https://sparontologies.github.io/deo/current/deo.html, accessed on 14 January 2024) which was written describing the major rhetorical elements of a document such as a journal article. Its classes include deo:Introduction, deo:Materials, deo:Methods, deo:Results, deo:RelatedWork, deo:FutureWork, etc. These rhetorical components give a defined rhetorical structure to the paper, which assists readers to identify the important aspects of the paper. DEO reuses some of the rhetorical blocks from the SALT Rhetorical Ontology [35] and extends them by introducing 24 additional classes. In the context of structural and rhetorical organization of scholarly articles, it was noted that the rhetoric organization of a paper does not necessarily correspond neatly to its structural components (sections, paragraphs, etc.). The Ontology of Rhetorical Blocks (orb) [36] introduces rhetorical classes to semantify sections of scholarly publications. Eg., orb:Introduction, orb:Methods, orb:Results, orb:Discussion to structure the Body of an article inspired after the IMRAD structure [37]. The hypothesis preceding this ontology is that the coarse rhetoric emerging from publications' content have commonly shared semantics. Thus ORB provided a minimal set of rhetorical blocks that could be leveraged from the Header, Body, and Tail of scholarly publications. The Ontology of Scientific Experiments [38], EXPO, advocated that the development of ontology of experiments–which are testbeds for cause-effect relations–is a fundamental step in the formalization of science. Reported scientific findings with their salient attributes buried in discourse is made explicit increasing the findability of problems with formal, semantic annotations supported by EXPO. It then constitutes the intermediate layer of a general ontology of scientific experiments with ontological concepts such as experimental goals, experimental methods and actions, types of experiments, rules for experimental design, etc., that are common between different scientific areas.

With the ontologies discussed, one observes that each ontology defines an information scope for formalization. The current level of formalization varies greatly in granularity and between the sciences. Ontology reuse [39] addresses in a sense the extent to which generalization or specification can occur depending on the level of the ontology model they are applied in, nonetheless is a key realizer in what would otherwise seem an impossible goal to design an ontology for Science. One of the first attempts to address the description of the whole publishing domain is the introduction of the Semantic Publishing and Referencing (SPAR) ontologies (http://www.sparontologies.net/, accessed on 14 January 2024). SPAR is a suite of orthogonal and complementary OWL2 ontologies that enable all aspects of the publishing process to be described in machine-readable metadata statements, encoded using RDF. It includes FaBiO, CiTO proposed by [40], BiRO, C4O proposed by [41], among others. Another noteworthy example that followed best practices in ontology development by reusing related ontologies [39] listed in the Linked Open Vocabularies (LOV) was Semsur, the Semantic Survey Ontology, proposed by [30,42]. It introduced the semantification model for survey articles as a core ontology for for describing individual research problems, approaches, implementations and evaluations in a structured, comparable way. It modeled metadata based on DCTerms, Semantic Web for Research Communities (SWRC) [43] and Friend of a Friend (FOAF) (http://xmlns.com/foaf/0.1/, accessed on 14 January 2024) ontologies. The inner structure of scientific articles was partially modeled by Discourse Elements Ontology (DEO) (http://www.sparontologies.net/ontologies/deo, accessed on 14 January 2024) and Linked Science Core (LSC) [44] to model publication workflows. Survey articles have been the traditional method for documenting overview of research progresses. However with the document-based publishing model, much of the data points presenting research progress remained buried in discourse, as a result were forever statically encoded. Semsur aimed to offer machine-actionability to these key resources.

2.2. Entity-Centric Annotation Models of Scholarly Publications

The trend towards scientific terminology mining methods in NLP steered the release of phrase-based annotated datasets in various domains. An early dataset in this line of

work was the ACL RD-TEC corpus [8] which identified seven conceptual classes for terms in the full-text of scholarly publications in Computational Linguistics, viz. *Technology and Method*; *Tool and Library*; *Language Resource*; *Language Resource Product*; *Models*; *Measures and Measurements*; and *Other*. Another dataset focused on the research dynamics discovery around scientific terminology in Computational Lingistics included the FTD corpus [7] annotated with *Focus*, *Task* and *Domain* of application entity types. Similar to terminology mining is the task of scientific keyphrase extraction. Extracting keyphrases is an important task in publishing platforms as they help recommend articles to readers, highlight missing citations to authors, identify potential reviewers for submissions, and analyse research trends over time. Scientific keyphrases, in particular, of type *Processes*, *Tasks* and *Materials* were the focus of the SemEval17 corpus annotations [10]. The dataset comprised annotations of the full text articles in Computer Science, Material Sciences, and Physics. Following suit was the SciERC corpus [12] of annotated abstracts from the Artificial Intelligence domain. It included annotations for six concepts, viz. *Task*, *Method*, *Metric*, *Material*, *Other-Scientific Term*, and *Generic*. Subsequently, based on this conceptual formalism, large-scale knowledge graphs such as AI-KG [14] and CS-KG [45] were generated. Recently, tackling the multidisciplinary discovery of entities, the STEM-ECR corpus [15] was introduced notably including the Science, Technology, Engineering, and Medicine domains. It was annotated with four generic concept types, viz. *Process*, *Method*, *Material*, and *Data* that mapped across all domains, and further with terms grounded in the real-world via Wikipedia/Wiktionary links. Furthermore, along the lines of the motivation of Agri-NER is the CS-NER service [18,19] that addresses the extraction of seven contribution-centric entities applicable in the Computer Science research field, viz. *Research problem*, *Resource*, *Method*, *Tool*, *Dataset*, *Language*, and *Solution* entity types from Computer Science paper titles and abstracts. These seven entity types were proposed to foster the discovery of research-domain-specific contribution templates in Computer Science.

Leaderboards construct of progress trackers taken up for the recording of results in the field of empirical Artificial Intelligence (AI) at large is a case in point of the development of templates arising from contribution-centric entities. This construct underlies the PapersWithCode https://paperswithcode.com/ (accessed on 14 January 2024) framework, as well as the ORKG Benchmarks https://orkg.org/benchmarks (accessed on 14 January 2024) feature. The construct defines the recording of results around four entity types viz. *Task*, *Dataset*, *Metric*, and *Score* from the full text of scholarly articles. The entities were then combined within the full-fledged semantic construct of a Leaderboard with between three or all four types for machine learning [13,17,46–48].

The Agri-NER service is situated within this latter broad paradigm of obtaining structured comparable, FAIR descriptions of scholarly contributions for the agricultural domain with the aim of bottom-up discovery of template patterns. However, it also relies on the first paradigm of scholarly knowledge structuring by mapping the automatically extracted terms to the AGROVOC ontology [49] which offers a controlled vocabulary designed to cover unambiguous semantic descriptions for terminology under the FAO's areas of interest. The following desiderata guided the creation of Agri-NER. (1) Manual curation of Agriculture named entities from 5500 article titles that reflect the contribution of a work enabling machine learning model training and development. (2) Associating terms within the AGROVOC ontology allowing for conceptual enrichment for the terms. (3) Allowing for ongoing, collaborative expert curation of named entities termwise and for their typing. (4) Juxtaposing a contribution-centric information extraction objective with term standardization in ontologies - why a simple term normalization against authoritative ontologies does not serve the objective of obtaining contribution-centric models? The rest of paper discusses how these requirements were accomplished.

In essence, our work's focus on FAIR principles, advanced NLP techniques, and the integration of machine-actionable knowledge capture aligns well with the core tenets of Industry 5.0, specifically in terms of its influence on agriculture [50,51]. Industry 5.0 emphasizes personalized and sustainable solutions, blending human-centric approaches

with advanced technological innovations [52]. The focus of the ORKG Agri-NER service, proposed in this work, on creating interoperable, reusable, and machine-interpretable models of scholarly contributions in agriculture fits into this paradigm by enabling more nuanced, efficient, and collaborative research practices. This approach can lead to more tailored agricultural practices and innovations, reflecting the personalized and sustainable ethos of Industry 5.0.

3. Materials and Methods

This section details the research methodology followed to develop the ORKG Agri-NER service. Our approach aligns with the typical methods used in Computer Science (CS) research within the field of NLP. This paper specifically addresses the development of a machine learning system for the automatic extraction of agricultural research-relevant entity types from scholarly articles, contributing to the ORKG database. The methodology adopted here is empirical and consists of four primary steps. The initial step involved defining the Agri-NER objective by selecting appropriate entity types pertinent to agriculture, based on theoretical considerations of various entity source types. This process is thoroughly explained in Section 3.1. Subsequently, from the theoretically identified entities, a set of seven specific entity types were chosen to formalize the ORKG Agri-NER goal, which we then defined. The chosen entity types and their definitions are detailed in Section 3.2.1. The third step entailed creating a human-annotated, gold-standard corpus of paper titles tagged with these seven entity types. Our aim was to build a corpus of high quality and size to facilitate the development of effective and robust machine learning models, as elaborated in Section 3.2.2. The final step involved training a machine learning system using this annotated corpus, which is discussed comprehensively in the dedicated section, Section 4.

3.1. Theoretical Paradigm: The AGROVOC Ontology and the ORKG Agri-NER Model

The work discussed in this paper seeks to integrate two paradigms of information representation: ontologized knowledge indexing supported by the AGROVOC ontology [49], and contribution-centric entity-based knowledge extraction supported by the ORKG Agri-NER model. The latter is the focus of this work. In both projects, the source material is taken from scholarly publications. The domain in both contexts is Agriculture where it is known that AGROVOC domain coverage is an amalgamation of the following related domains, viz. Agriculture, Fisheries, Forestry, and Environment. Detailed information about the respective projects i.e., AGROVOC [53–56] and ORKG [5,57] can be obtained elsewhere.

In a system like the ORKG (https://orkg.org/, accessed on 14 January 2024), it is impossible to exhaustively predict in advance which entity types will be needed to semantically model scholarly contributions in the vast domain of Agriculture. The list of entity types, for instance, is in principle open-ended for the main reason that scholarly innovations are continuously made. However, this implication also holds true for the AGROVOC ontology since, as research progresses, the existing agricultural terminology is constantly evolving, on the one hand, and new concepts are constantly discovered, on the other hand. In the context of the ORKG, our initial hypothesis is that starting out with an initial set of candidate entity types in Agri-NER as recommendations offer researchers a rough sketch to design templates aggregating one or more of the suggested entity types and define new types in addition to standardize the process of describing innovations across research papers addressing the same research problem or in the same domain, for instance. Given this, the workflow of Agri-NER will be constantly evolving as new entity types introduced by researchers describing their contributions will be periodically reviewed and fed back as input to retrain the models. In a sense, the evolution of AGROVOC is indeed based on the same principles.

Finally, we note that the notion of entity types from Agri-NER and concepts in AGROVOC are not equivalent. A concept in AGROVOC pertains to a real-world entity with alternate names. E.g., Maize https://agrovoc.fao.org/browse/agrovoc/en/page/c_12332

(accessed on 14 January 2024) is a concept in AGROVOC with alternate names such as "corn". In contrast, entity types in ORKG Agri-NER refer to the functional role of various real-world entities in the context of the contribution of a work which depends on the publication sentence discourse describing the contribution.

3.2. The ORKG Agri-NER Specifications

In this subsection, we first offer definitions for each of the seven entity types considered in ORKG Agri-NER. Following which, we discuss the annotation process and conclude with corpus statistics.

3.2.1. The Seven ORKG Agri-NER Entity Type Definitions

The Agri-NER model is structured around the following seven core entity types.

- RESEARCH PROBLEM. It is a natural language mention phrase of the theme of the investigation in a scholarly article [42]. Alternatively, in the Computer Science domain it is referred to as *task* [17] or *focus* [7]. An article can address one or more research problems. E.g., seed germination, humoral immunity in cattle, sunbird pollination, seasonal and inter-annual soil CO_2 efflux, etc. Generally, RESEARCH PROBLEM mentions are often found in the article Title, Abstract, or Introduction in the context of discourse discussions on the theme of the study; otherwise in the Results section in the context of findings discussions on the theme of the study.
- RESOURCE. They are either man-made or naturally occurring tangible objects that are directly utilized in the process of a research investigation as material to facilitate a study's research findings. "Resources are things that are used during a production process or that are required to cover human needs in everyday life" [53]. E.g., RESOURCE 'pesticides' used to study the RESEARCH PROBLEM 'survival of pines'; the PROCESS 'repeated migrations' studied over RESOURCE 'southern African members of the genus Zygophyllum'; RESOURCE 'Soil aggregate-associated heavy metals' studied in LOCATION 'subtropical China'. Resources are used to either address the RESEARCH PROBLEM or to obtain the SOLUTION.
- PROCESS. It is defined as an event with a continuous time-frame that is pertinent with a specific function or role to the theme of a particular investigation or research study. As defined in the AGROVOC ontology [53], a PROCESS can be a set of interrelated or interacting activities which transforms inputs into outputs, or simply a naturally occurring phenomenon that is studied. E.g., irradiance, environmental gradient, seasonal variation, quality control, salt and alkali stresses, etc.
- LOCATION. Includes all geographical locations in the world seen similar to the AGROVOC location concept [49] as a 'point in space'. Often LOCATION, in terms of relevance to the research theme, is the place where the study is conducted or a place studied for its RESOURCE or PROCESSes w.r.t. a RESEARCH PROBLEM. LOCATION mentions can be as fine-grained as having regional boundaries or as broad as having continental boundaries. E.g., Cape Floristic Region of South Africa, winter rainfall area of South Africa, sahel zone of Niger, southern continents, etc.
- METHOD. This concept imported from the Computer Science domain pertains to existing protocols used to support the solution [19]. The interpretation or definition of the concept similarly holds for the agricultural domain. It is a predetermined way of accomplishing an objective in terms of prespecified set of steps. E.g., On-farm comparison, semi-stochastic models, burrows pond rearing system, bradyrhizobium inoculation, electronic olfaction, systematic studies, etc.
- SOLUTION. It is a phrasal succinct mention of the novel contribution or discovery of a work that solves the RESEARCH PROBLEM [19]. The SOLUTION entity type is characterized by a long-tailed distribution of mentions determined by the new research discoveries made. The SOLUTION of one work can be used as a METHOD or TECHNOLOGY or comparative baselines in subsequent research works. Of all the entity types introduced in this work, SOLUTION like RESEARCH PROBLEM is specifically tailored

to the ORKG contribution model. E.g., radiation-induced genome alterations, artificially assembled seed dispersal system, commercial craftwork, integrated ecological modeling system, the MiLA tool, next generation crop models etc.

- TECHNOLOGY. Practical systems realized as tools, machinery or equipment based on the systematized application of reproducible scientific knowledge to reach a specifiable, repeatable goal. In the context of the agriculture domain, the goals would pertain to agricultural and food systems. E.g., stream and riverine ecosystem services, hyperspectral imaging, biotechnology, continuous vibrating conveyor, low exchange water recirculating aquaculture systems, etc.

3.2.2. The ORKG Agri-NER Corpus Annotation Methodology

Having introduced the seven contribution-centric entity types used in ORKG Agri-NER, we now elicit the methodology for producing the instance annotations for the entity types from a corpus of paper titles.

Raw Dataset

The first step entailed downloading a raw corpus comprising paper titles of scholarly articles published in the agricultural domain. For this, a sample size needed to be defined. In this regard, the corpus size needed to satisfy two criteria: a large enough sample size to train a machine learning model and a small enough sample size such that the human annotation task was feasible. As such after discussions with the human annotator a sample size of 5500 titles was arrived at. A corpus with thousands of data points easily satisfies the objective of obtaining a robust machine learning system. This we concur based on our prior work with training NER machine learning systems in a multidisciplinary setting [15,58] and a single domain setting [19]. Thus 5500 articles in text format and restricted only to the articles with the CC-BY redistributable license on Elsevier were first downloaded using the following list https://github.com/jd-coderepos/stem-ner-60k/blob/main/raw-data/Elsevier-ccby-articles-w-domain-mapping.tsv (accessed on 14 January 2024). Next, our aim was to obtain the seven entity type annotations for only the *titles* in this corpus of publications. For this, a raw dataset of the article titles was created https://github.com/jd-coderepos/contributions-ner-agri/tree/main/raw-data (accessed on 14 January 2024).

Corpus Annotation

With a corpus of titles in place, we were then first and foremost faced with a blank slate of entity types to annotate since there was no reported prior work for NER in the agricultural domain. A natural question here is how did we arrive at the seven entity types, viz. RESEARCH PROBLEM, RESOURCE, PROCESS, LOCATION, METHOD, SOLUTION, and TECHNOLOGY, defined earlier? This was done based on the following 3-step methodology.

1. A list of entity types used in our prior work [19] on *contribution-centric* NER for the Computer Science (CS) domain was created as a reference list. This list included the following CS-domain-specific contribution-centric types, viz. SOLUTION, RESEARCH PROBLEM, METHOD, RESOURCE, TOOL, LANGUAGE, and DATASET. We identified this as a suitable first step owing to the strong overlap of the annotation aim between our prior work on the CS domain and our present work on the agriculture domain, i.e., that of identifying *contribution-centric* entities from paper titles. We hypothesized that some entity types, e.g., RESEARCH PROBLEM, that satisfy the functional role of reflecting the contribution of scholarly articles by nature of their genericity could be applicable across domains. As such the listed CS-domain contribution-centric entity types were tested for this hypothesis. Furthermore, based on the successful annotation outcomes of paper titles offering a rich store of contribution-centric entities, this work focusing on a new domain, i.e., agriculture, similarly based its entity annotation task on paper titles. Thus, with an initial set of entities in place, our task was then to

identify the entities that were generic enough to be transferred from the CS domain to the domain of agriculture.

2. Considering that some new agriculture domain-specific entity types would also need to be introduced, a list of the 24 top-level concepts in the AGROVOC ontology [53] as the reference standard was drawn up. This list included concepts such as FEATURES (https://agrovoc.fao.org/browse/agrovoc/en/page/c_331061, accessed on 14 January 2024), LOCATION (https://agrovoc.fao.org/browse/agrovoc/en/page/c_330988, accessed on 14 January 2024), MEASURE (https://agrovoc.fao.org/browse/agrovoc/en/page/c_330493, accessed on 14 January 2024), PROPERTIES (https://agrovoc.fao.org/browse/agrovoc/en/page/c_49874, accessed on 14 January 2024), STRATEGIES (https://agrovoc.fao.org/browse/agrovoc/en/page/c_330991, accessed on 14 January 2024), etc. The focus was maintained only on the top-level concepts, since traversing lower levels in the ontology led to specific terminology defined as a concept space such as Maize https://agrovoc.fao.org/browse/agrovoc/en/page/c_12332 (accessed on 14 January 2024). Since specific terminology do not serve the purpose of reflecting a functional role, hence by their inherent nature were ruled out as conceptual candidates for contribution-centric entity types.
3. Given the two reference lists of generic CS domain entity types and domain-specific AGROVOC concepts from steps 1 and 2, respectively, the third step involved selecting and pruning the lists to arrive at a final set of contribution-centric entity types to annotate agricultural domain paper titles with. There were two prerequisites defined for arriving at the final set of entity types: (a) it needed to include as many of the generic entities as were semantically applicable; and (b) introduce new domain-specific types complementing the semantic interpretation of the generic types such that the final set could be used as a unit for contribution-centric entity recognition. Concretely, these requisites were realized as a pilot annotation task over a set of 50 paper titles performed by a postdoctoral researcher. Starting with the CS domain inspired list of generic entities, the pilot annotation task showed that the CS domain TOOL, LANGUAGE, and DATASET types were not applicable to agricultural domain. This left a set of four types, viz. SOLUTION, RESEARCH PROBLEM, METHOD, and RESOURCE for the final annotation task. For the domain-specific entities, via the pilot annotation exercise, it was fairly straightforward to prune out most of the AGROVOC concepts on the basis of the following three criteria. (a) Six concepts did not fit in the criteria of offering a functional role that reflected the contribution of a work. These were *entities*, *factors*, *groups*, *properties*, *stages*, *state*. (b) Nine concepts indicated that they were more paper content-specific than title-specific. These were *activities*, *events*, *features*, *measure*, *phenomena*, *products*, *site*, *systems*, and *time*. And, (c) since our objective was to capture the most generic entity satisfying the functional role of reflecting the paper contribution, some of the top-level AGROVOC concepts could be subsumed by others. Specifically, the four types viz. *objects*, *organisms*, *subjects*, *substances* were subsumed as AGROVOC *resources*. Also *strategies* was subsumed as AGROVOC *methods*. In the end, from an initial list of 25 types, pruning out 15 types and subsuming 5 types, we were left with a set of five types for the final annotation task, viz. *location*, *methods*, *processes*, *resources*, *technology*. Then the generic and domain-specific lists were resolved as follows: SOLUTION and RESEARCH PROBLEM originating from the CS domain were retained as is for the agriculture domain; AGROVOC *methods* was resolved to the generic METHOD type and AGROVOC *resources* was resolved to RESOURCE; the remaining AGROVOC entities were first lemmatized for plurals (e.g., *processes* → PROCESS) and otherwise retained as is for LOCATION and TECHNOLOGY types.

With the final list of seven contribution-centric entity types arrived at for the agricultural domain, the raw dataset of 5500 paper titles could then be annotated. Note that among the seven entity candidates, three or four entity types applied at most for annotating a paper title for its entities with a possibility for repeated occurrences of one or more types. To offer the reader an insightful look into our corpus, Table 2 illustrates with the help of color codes for the entity types, five annotated paper title instances as examples. To facilitate further research on this topic, our corpus is publicly released with the CC BY-SA 4.0 license at https://github.com/jd-coderepos/contributions-ner-agri (accessed on 14 January 2024).

Table 2. Six example instances in the ORKG Agri-NER corpus of annotated paper titles with the seven contribution-centric entity types, viz. RESOURCE, RESEARCH PROBLEM, PROCESS, LOCATION, METHOD, SOLUTION, and TECHNOLOGY.

Annotated Paper Titles
PICS bags safely store unshelled and shelled groundnuts in Niger TECHNOLOGY: PICS bags RESOURCE: unshelled and shelled groundnuts LOCATION: Niger
On-farm comparison of different postharvest storage technologies in a maize farming system of Tanzania Central Corridor METHOD: On-farm comparison RESEARCH PROBLEM: postharvest storage technologies TECHNOLOGY: maize farming system LOCATION: Tanzania Central Corridor
Comparing pressures on national parks in Ghana and Tanzania: The case of Mole and Tarangire National Parks RESEARCH PROBLEM: Comparing pressures on national parks LOCATION: Ghana, Tanzania, Mole and Tarangire National Parks
Ecological connectivity across ocean depths: Implications for protected area design RESEARCH PROBLEM: Ecological connectivity RESOURCE: ocean depths SOLUTION: protected area design
The truth about cats and dogs: Landscape composition and human occupation mediate the distribution and potential impact of non-native carnivores PROCESS: Landscape composition, human occupation RESEARCH PROBLEM: distribution and potential impact of non-native carnivores
Potential of metal contamination to affect the food safety of seaweed (*Caulerpa* spp.) cultured in coastal ponds in Sulawesi, Indonesia RESEARCH PROBLEM: metal contamination to affect the food safety RESOURCE: seaweed (*Caulerpa* spp.) LOCATION: coastal ponds in Sulawesi, Indonesia

The Agri-NER Corpus Statistics

Our corpus characteristics are further examined in terms of the overall corpus statistics shown in Table 3. From a raw dataset of 5500 paper titles, a total of 15,261 entity annotations were obtained with 10,406 of them being unique. The annotation rate in terms of number of entities annotated per title is at 2.93 entities. Of the 5500 titles, eight could not be annotated with any of the seven contribution-centric entity types. Hence the minimum number of entities/title shows a 0 count statistic. These eight titles were outliers in our corpus. Consider the two-token title "Garden Masterclass" as one example among the eight unannotatable titles of which the others similarly reflected a peculiar characteristic such as, for instance, being too short.

Table 3. Overall statistics of the gold-standard Open Research Knowledge Graph Agriculture Named Entity Recognition (ORKG Agri-NER) corpus.

Statistic Parameter	Counts
Num. Title Tokens overall	71,632
Max., Min., Avg. Num. Tokens/Title	65, 2, 13.75
Num. Entity Tokens overall	47,608
Max., Min., Avg. Num. Tokens/Entity	15, 1, 3.12
Num. Entities	15,261
Num. Unique Entities	10,406
Max., Min., Avg. Num. Entities/Title	9, 0, 2.93

Corpus statistics in terms of instantiated entities per entity type are shown in Table 4. We see that among the seven entity types, RESOURCE and RESEARCH PROBLEM are highly predominant as contribution-centric entity annotations.

Table 4. Statistics of the gold-standard Open Research Knowledge Graph Agriculture Named Entity Recognition (ORKG Agri-NER) corpus per the seven entity types annotated. The parenthesized numbers represent the unique entity counts.

Entity Type	Counts
Num. RESOURCE	5490 (4073)
Num. RESEARCH PROBLEM	4707 (3403)
Num. PROCESS	1789 (1525)
Num. LOCATION	1525 (776)
Num. METHOD	1364 (940)
Num. SOLUTION	250 (221)
Num. TECHNOLOGY	136 (113)

4. Results

With an annotated corpus in place, various neural machine learning models were evaluated to create the ORKG Agri-NER service. This section is devoted to discussions about our machine learning experimental setup and results from the various trained models to obtain an optimal ORKG Agri-NER automated service.

4.1. Experimental Setup

4.1.1. Dataset

The ORKG Agri-NER corpus presents a sequence labeling scenario. For learning a sequence labeler, each sentence is tokenized as a set of words where each word is assigned a classification symbol. The series of classification decisions over the words are then aggregated in a final step to extract classifications for phrases. Thus, in a first step, our raw annotated data had to be converted into a suitable format for machine learning. The most common representation format adopted for sequence labeling is called the CONLL format introduced in the CONLL 2003 shared task series [59]. Per the prescribed format, each line in the data file consists of tab-separated values with the tokenized word to be classified in the first column, features such as the part-of-speech (POS) tag in the columns in between, and the classification token in the last column. Sequences of tokenized words constitute consecutive lines in the data file. And an empty line separates sentence sequences. To create our data in this format, for tokenization the titles were simply split on spaces. In addition since we were interested in testing additional features as informative to the task or not, we obtained POS tags and NER tags for the tokens with the help of the Stanford Stanza library [60]. These features constituted the second and the third columns of our data file. Finally, the fourth column constituted the classification tag. For this we experimented with two well-known formats, viz. IOB and IOBES. The IOB tagging sequence [61] is the one

where the B- tag is used in the beginning of every phrasal entity type, I- prefix before a tag indicates that the tag is inside a phrasal entity type, and O tag indicates that a token belongs to no entity type. E.g., if the a phrase is of type METHOD, the tag for the first token of the phrase will be B-METHOD and all the remaining tokens of the phrase will be tagged I-METHOD. On the other hand, the IOBES tagging sequence [62] is the one with the tags B, E, I, S or O where S is used to represent a chunk containing a single token. Chunks of length greater than or equal to two always start with the B tag and end with the E tag.

Once our data was converted to the CONLL format, the annotated gold-standard collection of 5500 annotated titles was randomly split as 5000 titles in the training set, 200 titles in the development set for the tuning of hyperparameters of the machine learning models, and 300 titles in the test set. The resulting dataset is also part of the community release and can be accessed here https://github.com/jd-coderepos/contributions-ner-agri/tree/main/NCRFpp-input-format (accessed on 14 January 2024).

4.1.2. Models

In this age of the "deep learning tsunami" [63], neural sequence labeling models are the state-of-the-art technique. The neural models completely alleviated the traditional method of manual feature engineering. Instead in neural models, features are extracted automatically through network structures including long short-term memory (LSTM) [64] and convolution neural network (CNN) [65]. As such various network architectures have evolved with each class of models outperforming the others. One class of models belongs to word-level neural networks [66] where words of a sentence are given as input to a Recurrent Neural Network (RNN), specifically, an LSTM and each word is represented by its word embedding. Another class of models belongs to character-level neural networks [67] where a sentence is taken to be a sequence of characters. This sequence is passed through a CNN, predicting labels for each character. Character labels are transformed into word labels via post processing. The third and most successful class of models belongs to a combination of word+character neural networks [68,69] where the first layer represents words as a combination of a word embedding and a convolution over the characters of the word, following this with a Bi-LSTM layer over the word representations of a sentence.

Thus inspired from state-of-the-art neural sequence labelers [68–71], we leveraged the outperforming architectural variant, i.e., the "Char CNN + Word BiLSTM + CRF" neural sequence labeling model architecture. The model has three layers. *1. Character Sequence Layer* which relies on CNN neural encoders for character sequence information. Specifically, the sliding window approach captures local features, which are then max-pooled to obtain an aggregated encoding of the character sequence. *2. Word Sequence Layer* which relies on bidirectional LSTMs as the word sequence extractor. Since word contexts are a crucial feature to build optimal sequence labelers, the bidirectional LSTMs are shown to be most effective since they encode both the left and right context information of each word. The hidden vectors for both directions on each word are concatenated to represent the corresponding word. Further, the word representations were computed one of two ways: either directly from the data, or as precomputed vectorized embedding representations. We used GloVe embeddings [72]. And *3. Inference Layer* as the last layer for token classification by taking the extracted word sequence representations as features and assigning labels to the word sequence. In this layer, we leverage Conditional Random Fields (CRFs). Since CRFs are able to capture label dependencies in the output layer which leads to better predictions, their usage has resulted in many state-of-the-art neural sequence labeling models [69,71,73]. For implementation purposes, we leveraged the open-source toolkit called NCRF++ [74] (https://github.com/jiesutd/NCRFpp, accessed on 14 January 2024) based on PyTorch. Our experimental configuration files for model hyperparameter details including learning rate, dropout rate, number of layers, hidden size etc., are released as config files here https://gitlab.com/TIBHannover/orkg/nlp/experiments/orkg-agriculture-ner/ (accessed on 14 January 2024).

Aside from experimenting with different neural architectures, another class of models that have proven to be the state-of-the-art for sequence labeling are the transformer-based BERT language models [75]. These models are pretrained for language comprehension with a masked language modeling objective on a large-scale corpus comprising millions of articles and billions of tokens. As such there are variants of the pretrained transformer language models released. We test two model variants: the original BERT model trained on the BookCorpus [76] plus English Wikipedia; and a pretrained variant released over scientific text called SciBERT [77] trained on 1.14 M papers from Semantic Scholar [4] which consists of 18% papers from the computer science domain and 82% from the biomedical domain. The large-scale transformer language models obtained pretrained deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers. To obtain state-of-the-art models for downstream tasks, the pretrained model parameters are then finetuned via a task-specific architecture taking as input a task-specific dataset. For NER sequence labeling, the finetuning model consists of three components: (a) a token embedding layer comprising a per-sentence sequence of tokens, where each token is represented as a concatenation of BERT word embeddings and CNN-based character embeddings [68], (b) a token-level encoder with two stacked bidirectional LSTMs [64], and (c) a Conditional Random Field (CRF) based tag decoder [68]. Note the two features columns discussed earlier in the dataset section are not relevant for BERT models, thus can be removed from the data or replaced by dummy tokens. The dataset for BERT models is also released https://github.com/jd-coderepos/contributions-ner-agri/tree/main/BERT-input-format (accessed on 14 January 2024). For implementation purposes, we use the scikit-learn wrapper to finetune the two BERT variants based on the https://github.com/charles9n/bert-sklearn (accessed on 14 January 2024) package. Furthermore, we experiment with BERT-base-cased and SciBERT-base-cased pretrained models, respectively. The best model hyperparameters are released in our Jupyter notebook created for experimental purposes also available in our code repo.

Summarily, to investigate a state-of-the-art neural sequence labeler, we experiment with the "Char CNN + Word BiLSTM + CRF" neural architecture which were an early class of models offering best performances on sequence labeling tasks, where the word embeddings are computed directly on the training corpus or obtained from fixed word embedding models, e.g., GloVe. As a second class of models we experiment with a BERT-based transformer sequence labeler that obtains contextualized embeddings from a large-scale pretrained model and is finetuned on our downstream Agri-NER task based on our annotated corpus. A pictorial depiction of our end-to-end sequence labeling architecture is shown in Figure 4.

4.1.3. Evaluation Metrics

Evaluations are considered in two main settings: 1. strict, i.e., exact match; and 2. relaxed, i.e., inexact match where the gold answer is checked to be contained in the predicted answer. We elaborate on the relaxed match setting with an example. Given a title "Woody vegetation dynamics in a communally utilised semi-arid savanna in Bushbuckridge, South Africa" where "Woody vegetation dynamics" is annotated as *research problem*, if the machine predicts "vegetation dynamics", then this is marked as a true inexact match since two of the tokens in the gold-standard annotation are present in the prediction. In both settings, the standard Precision, Recall, and F1 score metrics are applied at the phrase level. Our phrase-based evaluation script can be accessed at https://github.com/jd-coderepos/contributions-ner-agri/blob/main/scripts/evaluate.py (accessed on 14 January 2024).

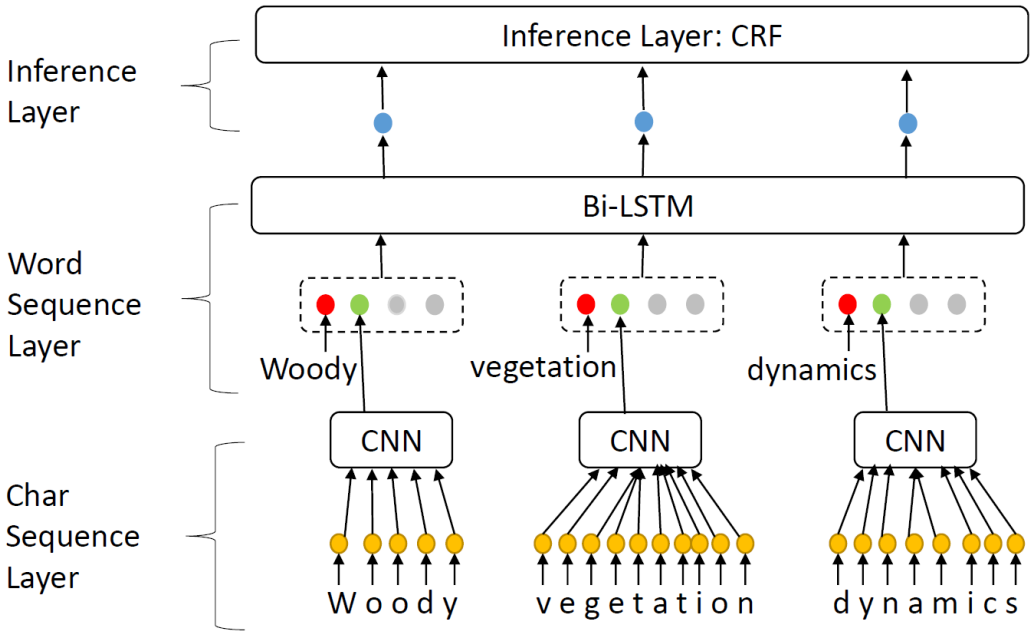

Figure 4. The traditional "Char CNN + Word BiLSTM + CRF" neural sequence labeling architecture for the input "Woody vegetation dynamics".

4.2. Experiments

In this section, we present the results and discuss observations from our two main sequence labeling strategies, respectively, and further contrast them w.r.t. each other. On the one hand, the "Char CNN + Word BiLSTM + CRF" sequence labeler resulted in 16 core experiments: one with no additional features, one with additional POS features, one with additional generic domain NER tag features, one with both POS and NER tags. Each of the four experiments were conducted in two scenarios: without and with GloVe embeddings. And each of the eight experiments were repeated in two tag encoding scenarios as IOBES and IOB tags. On the other hand, the BERT-based sequence labeler resulted in four total experiments: one with the BERT model variant and a second with the SciBERT model variant. And the two experiments repeated in the two tag encoding scenarios as IOBES and IOB tags. Thus overall 20 main experiments were conducted with additional sub-experiments within each category for model hyperparameter tuning.

The 16 core experiment results from the "Char CNN + Word BiLSTM + CRF" sequence labeler are reported in Table 5. And the four core experiment results from the transformer models are reported in Table 6. In the two tables, respectively, the best results for each of the precision, recall, and f-score metrics are highlighted in the bold, with the best F-scores overall in the exact versus inexact evaluation settings underlined. Next we discuss the experimental results with respect to five main research questions (RQ).

Table 5. Results from the state-of-the-art "Char CNN + Word BiLSTM + CRF" Neural Sequence Labeler. Numbers in bold are the highest score for each metric; numbers underlined are the overall highest exact and inexact match scores, respectively.

	IOBES						**IOB**					
	Exact Match			**Inexact Match**			**Exact Match**			**Inexact Match**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
no features	56.38	62.27	59.18	59.1	65.27	62.03	54.62	59.47	56.94	58.52	63.71	61.0
+GloVe	57.74	62.79	60.16	60.86	66.19	63.41	57.9	**63.49**	60.57	61.64	**67.59**	64.48
POS	57.11	61.88	59.40	60.24	65.27	62.66	56.0	60.53	58.18	60.42	65.3	62.76
+GloVe	57.5	63.58	60.38	60.09	66.45	63.11	56.63	63.11	59.7	60.45	67.38	63.73
NER	56.12	61.1	58.5	58.39	63.58	60.88	56.43	61.59	58.9	60.44	65.96	63.08
+GloVe	57.5	63.58	60.38	60.09	66.45	63.11	**58.32**	**63.49**	**60.8**	**62.09**	**67.59**	**64.72**
POS + NER	56.66	61.75	59.09	59.52	64.88	62.09	55.93	61.77	58.71	59.88	66.14	62.85
+GloVe	**58.23**	**63.71**	**60.85**	**60.98**	**66.71**	**63.72**	55.93	61.77	58.71	59.88	66.14	62.85

Table 6. Results from the state-of-the-art BERT-based Transformer Language Models. Numbers in bold are the highest score for each metric; numbers underlined are the overall highest exact and inexact match scores, respectively.

	IOBES						**IOB**					
	Exact Match			**Inexact Match**			**Exact Match**			**Inexact Match**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
BERT-Base-Cased	58.74	**66.71**	**62.47**	**62.64**	**71.15**	**66.63**	**60.58**	**68.25**	**64.19**	**64.09**	**72.2**	**67.91**
SciBERT-SciVocab-Cased	**58.78**	66.06	62.2	62.49	70.23	66.13	59.15	66.09	62.43	63.63	71.11	67.17

RQ1: How effective were the additional POS tag and generic domain NER tag features in the "Char CNN + Word BiLSTM + CRF" neural sequence labeler?

To answer this question, we examine the results reported in Table 5. Both tagging settings i.e., IOBES and IOB obtained improved scores with the additional features. On the one hand, the IOB tag representation experiments reported highest performances from NER tags. On the other hand, the IOBES tag representations, which constituted a larger classification space, benefited from the enriched feature representation space including both POS and the generic NER tags.

RQ2: Was initializing the word embeddings space with statically encoded embeddings from GloVe beneficial to the "Char CNN + Word BiLSTM + CRF" neural sequence labeler?

Contrasting the alternative rows in Table 5, we see that for each experimented feature setting, initialization of the word embeddings space with the precomputed GloVe embeddings obtained a better performing sequence labeler. Thus projecting the words in our dataset into an externally predefined semantic space formed from a larger external corpus was indeed more beneficial than computing words embeddings from the restricted space of the just the Agri-NER corpus.

RQ3: Which of the tag sequence representations, i.e., IOB versus IOBES, constituted the most effective task representation?

From the "Char CNN + Word BiLSTM + CRF" sequence labeler results reported in Table 5, the results were not conclusive. In the exact match settings, the IOBES tag sequence reported an insignificant 0.05% improvement with 60.85% F1 over the results from the IOB tag representation. In the inexact match settings, the IOB tag representation reported a 1% improvement with 64.72% F1 over the results from the IOBES tag representation. From the BERT-based sequence labeler results reported in Table 6, the results showed the IOB tag representation was the better format. In the exact match settings, the results

with the IOB representation was at 64.19% F1—2 points above the results with the IOBES representation at 62.47% F1. In the inexact match settings, again the results with the IOB tag representation was better at 67.91% F1—1 point above the results with the IOBES representation at 66.63% F1.

RQ4: Which method contrasting the results from the "Char CNN + Word BiLSTM + CRF" neural sequence labeler versus the BERT-based labeler produced the best results?

We examine the underlined results reported in Tables 5 and 6 from the "Char CNN + Word BiLSTM + CRF" sequence labeler and the BERT-based labeler, respectively. The BERT-based model significantly outperforms the "Char CNN + Word BiLSTM + CRF" in both settings including exact match with 64.19% F1 versus 60.85% and inexact match with 67.91% F1 versus 64.72% F1.

In light of the better performing BERT-based sequence labeler, revisiting RQ3, we claim that the IOB tag sequence representation is ideal given the ORKG Agri-NER corpus.

RQ5: Was a sequence labeler finetuned on a scholarly domain pretrained BERT variant more effective than a pretrained BERT variant on the generic domain?

Finally, comparing results between the scholarly domain SciBERT versus the generic domain BERT, we see that the generic domain BERT variant outperformed SciBERT. We can attribute these unexpected results observation on the fact that SciBERT is pretrained on data largely from the biomedical domain which is different from the agricultural domain. It remains to be explored in future work whether we can achieve boosted performances of our Agri-NER task given a large-scale pretrained model also covering agriculture.

Our source code is publicly released here https://gitlab.com/TIBHannover/orkg/nlp/experiments/orkg-agriculture-ner (accessed on 14 January 2024) with the MIT license. Based on the experimental results, the best model is released as the ORKG Agri-NER service available in two formats: 1) as a Python package at https://orkg-nlp-pypi.readthedocs.io/en/latest/services/services.html (accessed on 14 January 2024), and as a REST API that can be invoked directly online via the interactive documentation at https://orkg.org/nlp/api/docs#/annotation/annotates_agri_paper_annotation_agriner_post (accessed on 14 January 2024).

5. Discussion

> "The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web—a web of data that can be processed directly or indirectly by machines". [29]

The Web flourished based on the hypertext linked information principle. Hypertext linking of information on the Web as a global information space revolutionized information access by enabling users to traverse, search, share, and browse information with the all-pervasive technology of web browsers. With the formalization of the Semantic Web [29], these same principles that applied to information represented as document descriptions are being applied to data. This has fostered the evolution of the Web as a global information space of only linked documents to one where both documents and data are linked. A prerequisite to realizing the Semantic Web is what is called as establishing a Linked Open Data Cloud (LOD Cloud). Linked Data constitutes the LOD. In other words, the LOD Cloud is a KG that manifests as a Semantic Web of Linked Data via a small set of standardized technologies: URIs and HTTP as identification and access mechanism for data resources on the web, and RDF as content representation format. Thus Linked Data realizes the vision of evolving the Web into a global data commons as what is defined as the Semantic Web, allowing applications to operate on top of an unbounded set of data sources, via standardised access mechanisms [78]. The LOD Cloud https://lod-cloud.net/ (accessed on 14 January 2024) constitutes the central hub that allow users to start browsing in one open-access submitted data source and then navigate along links into related data sources. This global data space connects data from diverse domains such as geography,

government, life sciences, linguistics, media, scholarly publications, social networks etc. Without the Linked Data creation tools and technologies, earlier data creation processes always resulted in data silos worldwide with no access means of interaction or interoperability. Now, however, leveraging a small standardized set of technologies of the Linked Data creation paradigm, any data source can be submitted to the LOD Cloud fostering the building of the Semantic Web. In light of these technological inventions, the FAIR guiding principles [3] for scientific data creation can indeed be a practice.

The next natural question is, is the ORKG Agri-NER corpus released in the LOD Cloud? The response is *not yet*. However in this last concluding section of the paper, we set the stage for realizing the vision of releasing the ORKG Agri-NER corpus within the LOD Cloud to be taken up in future work. The research paradigms underlying the NLP production of data and the Semantic Web production of data over a new domain are particularly beset by several steps of methodological and technological considerations. This merits dedicated discussions of the respective paradigm research processes and outcomes. The NLP data production lifecycle focuses on instantiated data annotation and all the steps that precede it including selecting a task and defining a conceptual annotation space for the task. While the Semantic Web data production lifecycle focuses on data representation in a strict machine-readable semantic representation language such as RDF or OWL to facilitate axiomatic machine reasoning. In other words, it is a natural product of the following ingredients. (1) Open Standards—such as URI, URL, HTTP, HTML, RDF, RDF-Turtle (and other RDF Notations), the SPARQL Query Language, the SPARQL Protocol, and SPARQL Query Solution Document Types. And, (2) A modern DBMS platform—Virtuoso from OpenLink Software or Neo4J (https://neo4j.com/, accessed on 14 January 2024) as a graph database management system.

This work has described the NLP NER research paradigm over the novel agricultural domain. As such it entailed presenting the selected *contribution-centric* NER task for the agricultural domain, defining the selected entity types for annotation, and annotating a corpus of 5500 paper titles as instantiated data for Agri-NER. In following work, the aim is to address the Semantic Web research paradigm such that scholarly contribution resources in the agricultural domain will be made into FAIR and reusable Linked Data. Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets [78]. Machine-readability will utilize URIs and HTTP as identification and access mechanisms and RDF content representation. Meaning definition will be handled via a schema model. Links to external datasets will be handled as linking to the AGROVOC ontology [49] as it is the only other semantic representation model for the agricultural domain. As already alluded to, Agri-NER and AGROVOC prescribe different conceptual spaces for how the entities are expected to be processed by machines. Specifically, AGROVOC enables the processing of the entities within a terminologically defined semantic space. It provides concepts resolved to URIs and supplemented with RDF descriptions of thousands of terms in the FAO's area of interest. While ORKG Agri-NER permits the processing of the entities w.r.t. their functional role as reflecting the contribution of a scholarly work. By aiming to link the entities in our ORKG Agri-NER corpus to AGROVOC, we enable users to fetch an enriched representation of the terms such as: What is its terminological definition?, or What are the alternative term namings across languages?, or Which other data linkings can be facilitated via the Linked Data source in consideration? For instance, "Borneo" a LOCATION entity type from Agri-NER is first resolved to AGROVOC concept for Borneo as https://agrovoc.fao.org/browse/agrovoc/en/page/c_1017 (accessed on 14 January 2024). This Linked Data enriches the term with its definition, alternate names of Borneo in various languages, etc. Furthermore, the AGROVOC Linked Data connects to the DBpedia Linked Data source [79]. Thus via AGROVOC the concept Borneo is enriched via a DBpedia knowledge source link https://dbpedia.org/page/Borneo (accessed on 14 January 2024) which offers additional information such as its total geographical area, geo-coordinates, the total population size

etc. In this way, by adopting data linking the Linked Data principles will foster scaling the development approach of Agri-NER beyond a fixed, predefined data silo of capturing *contribution-centric* entities, to encompass a larger number of relevant structured knowledge sources on the LOD cloud comprising heterogeneous data models that each constitute unique semantic spaces for the machine-actionability of terms.

Toward FAIR, Reusable Scholarly Contributions in Agriculture, for machine readability and semantic representation, the schema and URI space will be implemented via global property and resource identifiers within the ORKG web ecosphere at https://orkg.org/ (accessed on 14 January 2024). And for obtaining Linked Data, AGROVOC will be utilized. In this section, we offer concrete implementation details that contrast ORKG Agri-NER and AGROVOC models as potential related Linked Data sources. The preliminary findings discussed in this paragraph are obtained w.r.t. the following research question. **RQ6:** *How many ORKG Agri-NER entities can be mapped to AGROVOC?* To answer the question, a programmatic process flow depicted in Figure 5 was established. The process was fairly straightforward. Given the terms annotated in the Agri-NER model, query the concept nodes in AGROVOC with the terms. For those terms that were found as a whole, the corresponding AGROVOC concept URI is the desired retrieval unit. For the terms that were not found as a whole, they were iteratively split as the longest spanning subphrases with subphrase lengths as: original phrase length $-$ 1 $\leq$ range $\leq$ 1. The link retrieval step was stopped when one or more of the subphrases for a specified subphrase length could be resolved to one or more AGROVOC concepts. Resultingly, some statistical insights shown in Table 7 were obtained. This will form the basis of Linked Data creation in future work toward realizing FAIR, Reusable Scholarly Contributions in Agriculture. Of all the entities annotated in Agri-NER, 16% of them are found as AGROVOC concepts. And 53.75% of the Agri-NER entities are found as subphrase AGROVOC concepts. Per Agri-NER entity type, the ones that were most linkable involved the least amount of subjectivity in phrasal boundary determination. One way of gauging the subjective boundary determination decisions for Agri-NER entity types from the least to most can be based on the proportion of the Agri-NER entity type terms that could be directly resolved to AGROVOC. From the least to the most, they were: LOCATION, TECHNOLOGY, PROCESS, METHOD, RESEARCH PROBLEM, RESOURCE, and SOLUTION. The corpus used in the analysis is publicly released https://github.com/jd-coderepos/contributions-ner-agri/tree/main/AGROVOC-linked-data-analysis (accessed on 19 January 2024).

Table 7. Statistics of the terms in the Open Research Knowledge Graph Agriculture Named Entity Recognition (ORKG Agri-NER) corpus that were linkable to the AGROVOC ontology overall (first three rows) and per the seven entity types annotated. The parenthesized numbers represent the proportion of entity phrases that could only be resolved to AGROVOC by one or more of their longest span subphrases.

Statistic Parameter	Counts
% Entities resolved (% Entities resolved as subphrases)	16.06% (53.75%)
Max., Min., Avg. phrase length resolved	5, 1, 1.55
Max., Min., Avg. subphrase length resolved	5, 1, 1.23
% LOCATION resolved (% LOCATION resolved as subphrases)	31.82% (41.57%)
% TECHNOLOGY resolved (% TECHNOLOGY RESOLVED AS SUBPHRASES)	17.99% (46.04%)
% PROCESS resolved (% PROCESS resolved as subphrases)	16.57% (52.22%)
% METHOD resolved (% METHOD resolved as subphrases)	15.11% (41.07%)
% RESEARCH PROBLEM resolved (% RESEARCH PROBLEM resolved as subphrases)	13.77% (60.5%)
% RESOURCE resolved (% RESOURCE resolved as subphrases)	13.68% (55.35%)
% SOLUTION resolved (% SOLUTION resolved as subphrases)	3.19% (60.56%)

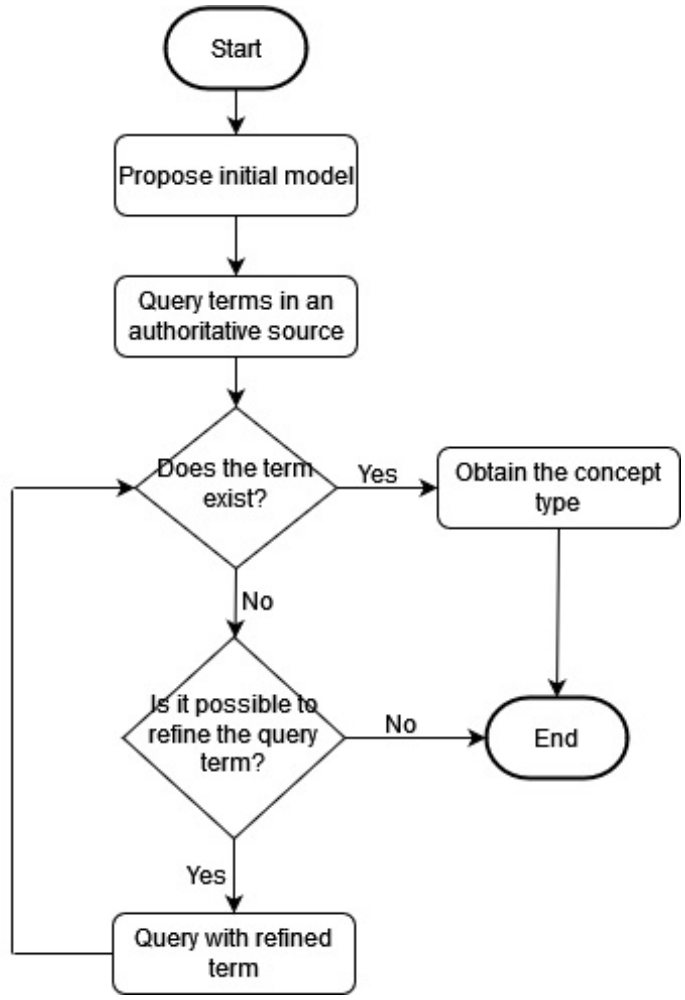

Figure 5. Process flow for linking Agri-NER entities to the AGROVOC ontology concept terms as an authoritative Linked Data source in the domain of Agriculture.

Future Directions

As we advance in the field of Agriculture NER, the integration and utilization of large language models (LLMs) present a promising avenue for future research and development [80]. These models, known for their deep learning capabilities and extensive training on diverse datasets, offer significant potential in enhancing the accuracy and scope of entity recognition in agricultural texts. The application of LLMs could revolutionize the way we extract, process, and interpret complex scientific entities, leading to more nuanced and contextually aware recognition systems. In context of furthering Agri-NER research, a key direction for future work is the customization of LLMs to better understand and interpret the unique terminologies and concepts specific to agriculture. This involves training models on domain-specific datasets such as ours, including scholarly articles and technical documents in the agricultural sector. Such specialized training would enable LLMs to accurately identify and classify a wide range of agricultural entities, thereby enhancing the overall quality and reliability of knowledge extraction in this field.

6. Conclusions

In this paper, we have introduced the Open Research Knowledge Graph Agriculture Named Entity Recognition (ORKG Agri-NER) corpus and service for contribution-centric scientific entity extraction and classification in the agricultural domain. The ORKG Agri-NER corpus is a benchmark for evaluating scientific entity extraction and classification in agriculture, using a generic conceptual formalism. This paper presents a baseline set of results on the benchmark leveraging state-of-the-art sequence labeling neural architectures

and transformer models. The paper also presents a 3-step automatic entity resolution procedure for mapping scientific entities to the AGROVOC ontology. The goal of this work is to provide a foundation for future research on automatic discovery of scientific entities in agricultural literature.

In conclusion, the development of the ORKG Agri-NER corpus and service represents a significant advancement in the field of agricultural NER. The utilization of machine-actionable representations and strategic reading techniques has demonstrated the potential to enhance the accessibility and interpretability of scholarly contributions in agriculture. The establishment of standardized entity types and the utilization of machine learning systems have shown promising results in the extraction and classification of scientific entities. Moving forward, the FAIR and reusable scholarly contributions in agriculture, facilitated by the ORKG Agri-NER service, hold the potential to significantly impact research, business, and organizational stakeholders within the agricultural domain.

Funding: Supported by TIB Leibniz Information Centre for Science and Technology, the EU H2020 ERC project ScienceGraph (GA ID: 819536) and the BMBF project SCINEXT (GA ID: 01lS22070).

Data Availability Statement: The dataset developed for this study can be found on the Github platform at https://github.com/jd-coderepos/contributions-ner-agri, accessed on 14 January 2024.

Acknowledgments: The author would like to acknowledge the ORKG Team for support with implementing the ORKG frontend interfaces of the Agriculture NER Annotator service.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

R&D	research and development
NLP	Natural Language Processing
NGO	non-govermental organization
FAIR	Findable, Accessible, Interoperable, and Reusable
KGs	Knowledge Graphs
ORKG	Open Research Knowledge Graph
IE	Information Extraction
NER	Named Entity Recognition
EPPO	European and Mediterranean Plant Protection Organization
Agri	Agriculture
DCTerms	Dublin Core Metadata Terms
DoCO	Document Components Ontology
DEO	Discourse Elements Ontology
ORB	Ontology of Rhetorical Blocks
EXPO	Ontology of Scientific Experiments
SPAR	Semantic Publishing and Referencing
LOV	Linked Open Vocabularies
SWRC	Semantic Web for Research Communities
FOAF	Friend of a Friend
LSC	Linked Science Core
AI	Artificial Intelligence
CS	Computer Science
POS	part-of-speech
LSTM	ong short-term memory
CNN	convolution neural network
RNN	Recurrent Neural Network

References

1. Johnson, R.; Watkinson, A.; Mabe, M. *The STM Report: An Overview of Scientific and Scholarly Publishing*; International Association of Scientific, Technical and Medical Publishers: Oxford, UK, 2018.
2. Renear, A.H.; Palmer, C.L. Strategic reading, ontologies, and the future of scientific publishing. *Science* **2009**, *325*, 828–832. [CrossRef] [PubMed]
3. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
4. Ammar, W.; Groeneveld, D.; Bhagavatula, C.; Beltagy, I.; Crawford, M.; Downey, D.; Dunkelberger, J.; Elgohary, A.; Feldman, S.; Ha, V.; et al. Construction of the Literature Graph in Semantic Scholar. In Proceedings of the NAACL-HLT, New Orleans, LA, USA, 1–6 June 2018; pp. 84–91.
5. Auer, S.; Oelen, A.; Haris, M.; Stocker, M.; D'Souza, J.; Farfar, K.E.; Vogt, L.; Prinz, M.; Wiens, V.; Jaradeh, M.Y. Improving access to scientific literature with knowledge graphs. *Bibl. Forsch. Prax.* **2020**, *44*, 516–529. [CrossRef]
6. Kim, S.N.; Medelyan, O.; Kan, M.Y.; Baldwin, T. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–16 July 2010; pp. 21–26.
7. Gupta, S.; Manning, C. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–13 November 2011; pp. 1–9.
8. QasemiZadeh, B.; Schumann, A.K. The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 1862–1868.
9. Moro, A.; Navigli, R. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; pp. 288–297.
10. Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; McCallum, A. SemEval 2017 Task 10: ScienceIE—Extracting Keyphrases and Relations from Scientific Publications. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 546–555. [CrossRef]
11. Gábor, K.; Buscaldi, D.; Schumann, A.K.; QasemiZadeh, B.; Zargayouna, H.; Charnois, T. Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In Proceedings of the 12th International Workshop on Semantic Evaluation, New Orleans, LA, USA, 5–6 June 2018; pp. 679–688.
12. Luan, Y.; He, L.; Ostendorf, M.; Hajishirzi, H. Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction. In Proceedings of the Conference Empirical Methods Natural Language Process (EMNLP), Brussels, Belgium, 31 October–4 November 2018.
13. Hou, Y.; Jochim, C.; Gleize, M.; Bonin, F.; Ganguly, D. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5203–5213. [CrossRef]
14. Dessì, D.; Osborne, F.; Reforgiato Recupero, D.; Buscaldi, D.; Motta, E.; Sack, H. Ai-kg: An automatically generated knowledge graph of artificial intelligence. In Proceedings of the International Semantic Web Conference, Online, 1–6 November 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 127–143.
15. D'Souza, J.; Hoppe, A.; Brack, A.; Jaradeh, M.Y.; Auer, S.; Ewerth, R. The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 2192–2203.
16. D'Souza, J.; Auer, S.; Pedersen, T. SemEval-2021 Task 11: NLPContributionGraph—Structuring Scholarly NLP Contributions for a Research Knowledge Graph. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Online, 5–6 August 2021; pp. 364–376. [CrossRef]
17. Kabongo, S.; D'Souza, J.; Auer, S. Automated Mining of Leaderboards for Empirical AI Research. In Proceedings of the International Conference on Asian Digital Libraries (ICADL 2021), Online, 1–3 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 453–470.
18. D'Souza, J.; Auer, S. Pattern-based acquisition of scientific entities from scholarly article titles. In Proceedings of the International Conference on Asian Digital Libraries (ICADL 2021), Online, 1–3 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 401–410.
19. D'Souza, J.; Auer, S. Computer science named entity recognition in the open research knowledge graph. In Proceedings of the International Conference on Asian Digital Libraries (ICADL 2022), Hanoi, Vietnam, 30 November–2 December 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 35–45.
20. SUNDHEIM, B. Overview of results of the MUC-6 evaluation. In Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, MA, USA, 6–8 November 1995.
21. Chinchor, N.; Robinson, P. MUC-7 named entity task definition. In Proceedings of the Seventh Conference on Message Understanding, Fairfax, VA, USA, 29 April–1 May 1998; Volume 29, pp. 1–21.
22. Sang, E.T.K.; De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, Edmonton, AB, Canada, 31 May–1 June 2003; pp. 142–147.

23. Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; Weischedel, R. OntoNotes: The 90% solution. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York, NY, USA, 5–7 June 2006; pp. 57–60.
24. Batbayar, E.E.T.; Tsogt-Ochir, S.; Oyumaa, M.; Ham, W.C.; Chong, K.T. Development of ISO 11783 Compliant Agricultural Systems: Experience Report. In *Automotive Systems and Software Engineering*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 197–223.
25. Oksanen, T.; Öhman, M.; Miettinen, M.; Visala, A. ISO 11783–Standard and its Implementation. In *IFAC Proceedings Volumes*; Elsevier: Amsterdam, The Netherlands, 2005; Volume 38, Number 1, pp. 69–74.
26. Le Bourgeois, T.; Marnotte, P.; Schwartz, M. The use of EPPO Codes in tropical weed science. In Proceedings of the EPPO Codes Users Meeting 5th Webinar, Online, 22 June 2021.
27. Shotton, D. Semantic publishing: The coming revolution in scientific journal publishing. *Learn. Publ.* **2009**, *22*, 85–94. [CrossRef]
28. Lis-Balchin, M.T. A chemotaxonomic reappraisal of the Section Ciconium Pelargonium (Geraniaceae). *S. Afr. J. Bot.* **1996**, *62*, 277–279. [CrossRef]
29. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
30. Fathalla, S.; Vahdati, S.; Auer, S.; Lange, C. SemSur: A core ontology for the semantic representation of research findings. *Procedia Comput. Sci.* **2018**, *137*, 151–162. [CrossRef]
31. Vogt, L.; D'Souza, J.; Stocker, M.; Auer, S. Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells. In Proceedings of the JCDL'20, Wuhan, China, 1–5 August 2020.
32. DCMI Usage Board. *Dublin Core Metadata Initiative Dublin Core Metadata Element Set*, Version 1.1; DCMI Usage Board: Metro Manila, Philippines, 2008.
33. Baker, T. Libraries, languages of description, and linked data: A Dublin Core perspective. *Library Hi Tech.* **2012**, *30*, 116–133. [CrossRef]
34. Constantin, A.; Peroni, S.; Pettifer, S.; Shotton, D.; Vitali, F. The document components ontology (DoCO). *Semant. Web.* **2016**, *7*, 167–181. [CrossRef]
35. Groza, T.; Handschuh, S.; Möller, K.; Decker, S. SALT-Semantically Annotated LaTeX for Scientific Publications. In Proceedings of the European Semantic Web Conference, Innsbruck, Austria, 3–7 June 2007; Springer: Berlin/Heidelberg, Germany, 2007 ; pp. 518–532.
36. Ciccarese, P.; Groza, T. Ontology of Rhetorical Blocks (orb). Editor's Draft. *World Wide Web Consortium*. 5 June 2011. Available online: http://www.w3.org/2001/sw/hcls/notes/orb/ (accessed on 12 May 2012).
37. Sollaci, L.B.; Pereira, M.G. The introduction, methods, results, and discussion (IMRAD) structure: A fifty-year survey. *J. Med. Libr. Assoc.* **2004**, *92*, 364.
38. Soldatova, L.N.; King, R.D. An ontology of scientific experiments. *J. R. Soc. Interface* **2006**, *3*, 795–803. [CrossRef] [PubMed]
39. Simperl, E. Reusing ontologies on the Semantic Web: A feasibility study. *Data Knowl. Eng.* **2009**, *68*, 905–925. [CrossRef]
40. Peroni, S.; Shotton, D. FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *J. Web Semant.* **2012**, *17*, 33–43. [CrossRef]
41. Di Iorio, A.; Nuzzolese, A.G.; Peroni, S.; Shotton, D.M.; Vitali, F. Describing bibliographic references in RDF. In Proceedings of the SePublica, Anissaras, Greece, 25 May 2014.
42. Fathalla, S.; Vahdati, S.; Auer, S.; Lange, C. Towards a knowledge graph representing research findings by semantifying survey articles. In Proceedings of the International Conference on Theory and Practice of Digital Libraries, Thessaloniki, Greece, 18–21 September 2017; Springer: Berlin/Heidelberg, Germany, 2017 ; pp. 315–327.
43. Sure, Y.; Bloehdorn, S.; Haase, P.; Hartmann, J.; Oberle, D. The SWRC Ontology–Semantic Web for Research Communities. In *Progress in Artificial Intelligence: 12th Portuguese Conference on Artificial Intelligence, EPIA 2005, Covilhã, Portugal, 5–8 December 2005;* Proceedings 12; Springer: Berlin/Heidelberg, Germany, 2005; pp. 218–231.
44. Baglatzi, A.; Kauppinen, T.; Keßler, C. Linked Science Core Vocabulary Specification. *Tech. Rep.* 2011. Available online: http://linkedscience.org/lsc/ns (accessed on 14 January 2024).
45. Dessí, D.; Osborne, F.; Reforgiato Recupero, D.; Buscaldi, D.; Motta, E. CS-KG: A Large-Scale Knowledge Graph of Research Entities and Claims in Computer Science. In Proceedings of the International Semantic Web Conference, Hangzhou, China, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022 ; pp. 678–696.
46. Jain, S.; van Zuylen, M.; Hajishirzi, H.; Beltagy, I. SciREX: A Challenge Dataset for Document-Level Information Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7506–7516. [CrossRef]
47. Mondal, I.; Hou, Y.; Jochim, C. End-to-End Construction of NLP Knowledge Graph. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Bangkok, Thailand, 1–6 August 2021; pp. 1885–1895. [CrossRef]
48. Kabongo, S.; D'Souza, J.; Auer, S. Zero-Shot Entailment of Leaderboards for Empirical AI Research. In Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Santa Fe, NM, USA, 26–30 June 2023; pp. 237–241. [CrossRef]
49. Subirats-Coll, I.; Kolshus, K.; Turbati, A.; Stellato, A.; Mietzsch, E.; Martini, D.; Zeng, M. AGROVOC: The linked data concept hub for food and agriculture. *Comput. Electron. Agric.* **2022**, *196*, 105965. [CrossRef]
50. Guruswamy, S.; Pojić, M.; Subramanian, J.; Mastilović, J.; Sarang, S.; Subbanagounder, A.; Stojanović, G.; Jeoti, V. Toward better food security using concepts from industry 5.0. *Sensors* **2022**, *22*, 8377. [CrossRef] [PubMed]

51. Baryshnikova, N.; Altukhov, P.; Naidenova, N.; Shkryabina, A. Ensuring global food security: Transforming approaches in the context of agriculture 5.0. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *988*, 032024. [CrossRef]
52. Akundi, A.; Euresti, D.; Luna, S.; Ankobiah, W.; Lopes, A.; Edinbarough, I. State of Industry 5.0—Analysis and identification of current research trends. *Appl. Syst. Innov.* **2022**, *5*, 27. [CrossRef]
53. AGROVOC Webpage. 2022. Available online: https://www.fao.org/agrovoc/home (accessed on 12 October 2022).
54. Soergel, D.; Lauser, B.; Liang, A.; Fisseha, F.; Keizer, J.; Katz, S. Reengineering thesauri for new applications: The AGROVOC example. *J. Digit. Inf.* **2004**, *4*, 1–23.
55. Lauser, B.; Sini, M.; Liang, A.; Keizer, J.; Katz, S. From AGROVOC to the Agricultural Ontology Service/Concept Server. An OWL model for creating ontologies in the agricultural domain. In Proceedings of the Dublin Core Conference Proceedings, Dublin Core DCMI, Manzanillo, Mexico, 3–6 October 2006.
56. Mietzsch, E.; Martini, D.; Kolshus, K.; Turbati, A.; Subirats, I. How Agricultural Digital Innovation Can Benefit from Semantics: The Case of the AGROVOC Multilingual Thesaurus. *Eng. Proc.* **2021**, *9*, 17.
57. Auer, S. Towards an Open Research Knowledge Graph. 2018. Available online: https://zenodo.org/records/1157185 (accessed on 14 January 2024).
58. Brack, A.; D'Souza, J.; Hoppe, A.; Auer, S.; Ewerth, R. Domain-independent extraction of scientific concepts from research articles. In Proceedings of the European Conference on Information Retrieval (ECIR 2020), Online, 14–17 April 2020; Springer: Berlin/Heidelberg, Germany, 2020 ; pp. 251–266.
59. Sang, E.F.T.K.; De Meulder, F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Development* **1837**, *922*, 1341.
60. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020.
61. Ramshaw, L.A.; Marcus, M.P. Text chunking using transformation-based learning. In *Natural Language Processing Using Very Large Corpora*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 157–176.
62. Krishnan, V.; Ganapathy, V. Named Entity Recognition. *Stanf. Lect. CS229.* 2005. Available online: http://cs229.stanford.edu/proj2005/KrishnanGanapathy-NamedEntityRecognition.pdf (accessed on 14 January 2024).
63. Manning, C.D. Computational linguistics and deep learning. *Comput. Linguist.* **2015**, *41*, 701–707. [CrossRef]
64. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
65. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
66. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
67. Kim, Y.; Jernite, Y.; Sontag, D.; Rush, A.M. Character-aware neural language models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
68. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1064–1074.
69. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural Architectures for Named Entity Recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 260–270.
70. Chiu, J.P.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
71. Peters, M.; Ammar, W.; Bhagavatula, C.; Power, R. Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1756–1765.
72. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]
73. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
74. Yang, J.; Zhang, Y. NCRF++: An Open-source Neural Sequence Labeling Toolkit. In Proceedings of the Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia, 15–20 July 2018; pp. 74–79.
75. Kenton, J.D.M.W.C.; Toutanova, L.K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
76. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 19–27.
77. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: Pretrained Language Model for Scientific Text. *arXiv* **2019**, arXiv:1903.10676.
78. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data: The story so far. In *Semantic Services, Interoperability and Web Applications: Emerging Concepts*; IGI Global: Hershey, PA, USA, 2011; pp. 205–227.

79. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
80. D'Souza, J. A Catalog of Transformer Models. 2023. Available online: https://orkg.org/comparison/R609337/ (accessed on 19 January 2024).