ISSA: Generic Pipeline, Knowledge Model and Visualization tools to Help Scientists Search and Make Sense of a Scientific Archive

Anne Toulet, Franck Michel, Anna Bobasheva, Aline Menin, Sébastien Dupré, Marie-Claude Deboin, Marco Winckler, Andon Tchechmedjiev

ISSA: Generic Pipeline, Knowledge Model and Visualization Tools to Help Scientists Search and Make Sense of a Scientific Archive

Anne Toulet[1], Franck Michel[2(✉)], Anna Bobasheva[3], Aline Menin[3], Sébastien Dupré[1], Marie-Claude Deboin[1], Marco Winckler[3], and Andon Tchechmedjiev[4]

[1] CIRAD (French Agricultural Research Centre for International Development), Ales, France
{anne.toulet,sebastien.dupre,marie-claude.deboin}@cirad.fr
[2] University Côte d'Azur, CNRS, Inria, Ales, France
franck.michel@inria.fr
[3] University Côte d'Azur, Inria, CNRS, Ales, France
{anna.bobasheva,aline.menin,marco.winckler}@inria.fr
[4] Euromov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
andon.tchechmedjiev@mines-ales.fr

Abstract. Faced with the ever-increasing number of scientific publications, researchers struggle to keep up, find and make sense of articles relevant to their own research. Scientific open archives play a central role in helping deal with this deluge, yet keyword-based search services often fail to grasp the richness of the semantic associations between articles. In this paper, we present the methods, tools and services implemented in the ISSA project to tackle these issues. The project aims to (1) provide a generic, reusable and extensible pipeline for the analysis and processing of articles of an open scientific archive, (2) translate the result into a semantic index stored and represented as an RDF knowledge graph; (3) develop innovative search and visualization services that leverage this index to allow researchers, decision makers or scientific information professionals to explore thematic association rules, networks of co-publications, articles with co-occurring topics, etc. To demonstrate the effectiveness of the solution, we also report on its deployment and user-driven customization for the needs of an institutional open archive of 110,000+ resources. Fully in line with the open science and FAIR dynamics, the presented work is available under an open license with all the accompanying documents necessary to facilitate its reuse. The knowledge graph produced on our use-case is compliant with common linked open data best practices.

Keywords: Data indexing · Scientific literature · Information retrieval · Linked open data · Knowledge graph

U. Sattler et al. (Eds.): ISWC 2022, LNCS 13489, pp. 660–677, 2022.
https://doi.org/10.1007/978-3-031-19433-7_38

1 Searching Scientific Literature: Beyond Keywords

In recent years, several evolutions have drastically transformed the way researchers interact with scientific literature. First, the number and pace of articles published are skyrocketing, such that it is increasingly difficult to keep up, find relevant articles or even identify potential collaborators. The use of social networks such as Twitter to monitor scientific advances, results in an echo chamber highlighting laboratories and researchers that are already visible and recognized. Second, most scientific literature repositories offer simple search capabilities that typically rely on keyword matches or author names. Such an approach commonly fails to grasp the richness of the semantic relationships that hold between articles, leaving to the user a cumbersome filtering of search results. Finally, the ultra-specialization of research communities makes it difficult to discover cross-disciplinary knowledge, yet essential to meet the growing demand of funding agencies for pluri- or interdisciplinarity. It is therefore essential to **offer tools that allow researchers, as well as scientific and technical information (STI) professionals, to find their way in and make sense of this mass of knowledge.** There exists a variety of methods and tools designed to process the content of text documents, extract knowledge, and provide advanced services. However, to the best of our knowledge, these tools are either domain-specific or address specific steps but do not provide an end-to-end, integrated pipeline.

In this paper, we present the methods, tools and services implemented in the ISSA project [3] to tackle these needs. ISSA aims to (1) provide **a generic, reusable and extensible pipeline for the analysis and processing of an open scientific archive**, (2) translate the results into a **semantic index in the form of an RDF knowledge graph** (KG); (3) develop **innovative search and visualization services exploiting the index**, aimed at researchers, decision makers, or STI professionals. Geared towards genericity and reusability, the proposed solution adheres to the FAIR principles [35] and the open science guidelines. Furthermore, ISSA adopts a pragmatic approach that strives to rely on robust, industry-proven, scalable solutions, and integrate them into a coherent, easily deployable pipeline.

The processing pipeline, depicted in Fig. 1, involves various artificial intelligence techniques: natural language processing, knowledge engineering, semantic web and linked data. Publications' metadata and full text are processed in order to extract thematic descriptors[1] and named entities (NE). To allow services to reason upon the extracted knowledge while leveraging terminological references such as ontologies or thesauri, thematic descriptors and NEs are linked with resources such as Wikidata, DBpedia and GeoNames. The resulting KG serves as a keystone able to support the development of services such as search and visualization. In particular, the Arviz [24] and MGExplorer [25] visualization tools make it possible to explore and visualize thematic association rules, networks

[1] Thematic descriptors are keywords linked to reference vocabularies, thesauri or ontologies that characterize an article as a whole. Unlike keywords provided by authors, they are extracted automatically using text classification methods.

of co-publications, or of articles with co-occurring topics, in order to concretely answer competency questions. These visualization tools are highly configurable and can be tailored to a wide range of scenarios.

To demonstrate the effectiveness of the proposed solution, we deployed it for the needs of a real-world use case, Agritrop [1], CIRAD[2]'s open archive of 110,000+ resources (i.e., book, book chapter, article, thesis, etc.). By drawing on the outcome of interviews conducted with CIRAD researchers and documentalists, we show the ability of these services to meet user needs and competency questions with relevant answers.

In the rest of this paper, Sect. 2 provides an overview and a comparison with related work. Section 3 describes the pipeline spanning metadata retrieval, extraction and linking of thematic descriptors and NEs, and construction of the KG. Then, Sect. 4 presents the exploitation and visualization tools and how they were configured in the Agritrop use-case. Section 5 provides further information about the accessibility of the pipeline and the KG generated in the case of Agritrop. Finally, Sect. 6 discusses the impact and reuse of this work in various communities, and Sect. 7 draws conclusions and suggests future works.

2 Related Works

For over twenty years, the open science movement has aimed at making scientific research results freely accessible, considerably transforming the landscape of scientific production. Initiatives such as **Research Data Alliance** [6] (RDA) that federates working groups on **FAIR principles**, metadata standards, and semantic resources (ontologies, thesauri, etc.); or **Go Fair** [2] and **European Open Science Cloud** (EOSC) [11], have laid the ground work for the implementation of the FAIR principles for open science. In this context, the role of open archives and of how to exploit them are central questions: many projects, including the ISSA project, have taken up this dimension, covering complementary aspects.

The **OpenMinted** [5] project aimed at creating a generic Software As A Service EU infrastructure for text mining, based on a modular architecture, that researchers could use by contributing their use-cases. After 5 years of development, the project fell short of delivering a fully functional prototype, merely laying the foundational components of the infrastructure. The related **Visa TM** project [7] was to be the core knowledge extraction component, integrating thesauri and ontologies from many domains, but only achieved a very preliminary integration [20]. In contrast, the ISSA project adopts a more modest but focused and pragmatic approach, proposing a generic pipeline adaptable to multiple domains, based on the integration of robust, industry-proven and scalable existing tools, and deployable by each community. ISSA also has a strong focus on using Linked Open Data and FAIR principles, which are absent from OpenMinted.

[2] CIRAD is the French Agricultural Research Centre for International Development https://www.cirad.fr/en.

The ISTEX infrastructure, which was meant to be the corpus provider for OpenMinted [20], has goals related to ISSA in that it aims at constituting corpora of scientific publications and providing research communities with tools to explore relevant subsets of the curated corpora. However, the main focus is to allow the creation and download of subsets of corpora through very precise criteria, extract terminology and provide a descriptive visualization of the results through the LODEX tool [10]. The indexing and consolidated KG aspects of ISSA are absent. The more recent **Covid-on-the-Web** project [28] has the most in common with ISSA, providing researchers with ways to access, extract and query knowledge from literature related to the coronavirus family, by building and exploiting a KG describing the concepts and arguments extracted from 100,000+ scientific articles, but stopping short of an end-to-end, reusable pipeline like in ISSA.

In summary the overall scope of ISSA includes something absent from all those initiatives: a generic end-to-end pipeline, that is easy to deploy and customize.

3 From an Open Scientific Archive to the ISSA Pipeline and Knowledge Graph

The ISSA pipeline harnesses existing tools to analyze and index the articles of a scientific archive, drawing meaningful links between the articles and the Web of Data, and following Semantic Web standards. Figure 1 describes the pipeline: (1) Metadata is retrieved from the open archive API, (2) translated into RDF with Morph-xR2RML and stored in a Virtuoso OS server. (3) Full text is extracted with Grobid and for each article, (4) Thematic descriptors and NEs are extracted from the text and linked to Wikidata, DBPedia and optionally domain-specific thesauri (unsupervised linking and disambiguation). (5) Descriptors and

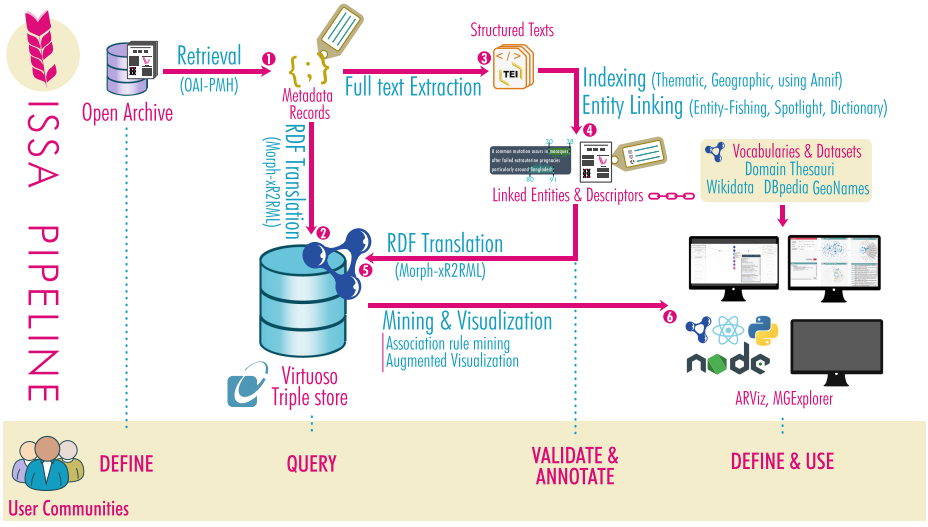

Fig. 1. ISSA pipeline: resources, services and applications.

entities are translated into a unified RDF dataset and stored in Virtuoso along metadata records. (6) The KG is exploited to propose augmented visualization applications.

3.1 Text Classification of Articles for Their Thematic Indexing

Thematic descriptors are keywords (typically 5 or 6) or expressions that characterize an article as a whole and that are linked to a standardized vocabulary. In some institutions, documentalists manually annotate articles with descriptors, which yields accurate annotations but is time consuming, such that it is usually not performed retroactively for older publications, possibly leaving behind a large set of legacy publications.

Provided that there exists a large enough corpus annotated with a domain vocabulary, one can train a specialized supervised classification model to automatically assign thematic descriptors to publications. The ISSA pipeline includes such a classification system through the integration of Annif [32], a framework developed by the National Library of Finland. Annif does not propose any new methods per se, but provides a framework and API to integrate existing machine learning models and tools to index corpora of scientific publications. In addition to the integration of multiple supervised and unsupervised models (TensorFlow deep net, Omikuji, fastText and Gensim), Annif supports multiple vocabulary formats, comes with standardized evaluation protocols and metrics, and supports multiple languages. In the ISSA pipeline, a corpus is extracted per language and split into training, validation and testing sets, in order to train the Annif model. The recreation of new models can be triggered independently from the pipeline, either manually or automatically at fixed intervals. The trained models are used in the pipeline to classify each article. For articles already manually indexed we end up with two sets of descriptors, one set corresponding to manual annotation and one set corresponding to automatic annotation.

Thematic descriptors are represented in RDF as annotations using the Web Annotation Vocabulary [34] (`issa:ThematicDescriptorAnnotation` is a subclass of `oa:Annotation`). An example is given in Listing 1.1 (lines 7–13). The annotation points to the annotated article (the target) and the resource that the descriptor links to (the body). It also provides the confidence of the extraction and linking of the descriptor, its rank in the list of descriptors ordered by descending confidence. Using PROV-O[3], the annotation keeps track of whether a thematic descriptor was retrieved from the article metadata or extracted by Annif.

Application to Agritrop. CIRAD curators annotate newly submitted articles with terms from AGROVOC [13], a standard SKOS thesaurus in the agronomy and agriculture domains. To train Annif to annotate new articles with AGROVOC terms, we extracted a corpus of approximately 12,000 English and French open-access articles. Descriptors manually annotated by curators were retrieved from Agritrop. For each language, separate training sets were created

[3] https://www.w3.org/TR/prov-o/.

```
@prefix dct:     <http://purl.org/dc/terms/> .
@prefix issa:    <http://data-issa.cirad.fr/> .
@prefix issapr:  <http://data-issa.cirad.fr/property/> .
@prefix oa:      <http://www.w3.org/ns/oa#> .
@prefix prov:    <http://www.w3.org/ns/prov#> .
@prefix schema:  <http://schema.org/> .
# Thematic descriptor "sustainable development"
[]  a                    prov:Entity , issa:ThematicDescriptorAnnotation ;
    issapr:confidence    0.4556 ;
    issapr:rank          6 ;
    oa:hasBody           <http://aims.fao.org/aos/agrovoc/c_35332> ;
    oa:hasTarget         <http://data-issa.cirad.fr/article/543654> ;
    prov:wasAttributedTo issa:AnnifSubjectIndexer .
# Named entity "banana"
[] a                 prov:Entity , oa:Annotation ;
   schema:about      <http://data-issa.cirad.fr/article/543654> ;
   issapr:confidence 0.5939 ;
   oa:hasBody        <http://www.wikidata.org/entity/Q503> ;
   oa:hasTarget [
     oa:hasSource    <http://data-issa.cirad.fr/article/543654#body_text> ;
     oa:hasSelector  [
       a oa:TextPositionSelector , oa:TextQuoteSelector ;
       oa:exact "banana" ; oa:start 12750; oa:end  12756 ]] ;
   prov:wasAttributedTo issa:EntityFishing .
```

Listing 1.1. Representation of a thematic descriptor extracted by Annif and linked to AGROVOC, and a named entity extracted from the article's body by Entity-fishing, and linked to Wikidata.

based on automatic language detection[4]. We experimented with different available models and chose the best performing one, namely an ensemble of lexical matching (MLLM) [32] and a tree-based machine learning algorithm [30].

3.2 Extraction and Linking of Named Entities

The ISSA pipeline relies on three tools to identify, disambiguate and link NEs from the articles (title, abstract and body) of the scientific archive:

- DBpedia Spotlight [15] annotates text in eight different languages with DBpedia entities. Disambiguation is carried out by entity linking using a generative model with maximum likelihood.
- Entity-fishing [31] identifies and disambiguates NEs against Wikidata. It relies on FastText word embeddings to generate candidates and ranks them with gradient tree boosting and features derived from relations and context.
- Dictionary projection annotation performs in-domain NEs with pyclinrec[5] and disambiguation is performed with EigenThemes [9] using hyperbolic graph embeddings [14] computed from the corresponding domain thesauri.

For each article, the pipeline invokes each of the three tools and translates their respective outputs into an RDF representation. An additional post-processing step specifically identifies geographic entities by looking for GeoNames mappings in the corresponding Wikidata concepts.

[4] https://pypi.org/project/pycld2/.
[5] https://github.com/twktheainur/pyclinrec.

Like thematic descriptors, NEs are modelled in RDF as annotations, as exemplified in Listing 1.1 (lines 14–23). The annotation points to the annotated article (property `schema:about`). The matched text fragment is described in the annotation target that points to the article part wherein the NE was recognized (title, abstract or body), and locates it with start and end offsets. The annotation body is the URI of the resource that the NE links to (Wikidata and Geonames in the example). The annotation includes the extraction and linking confidences, and provenance information regarding the tool used to extract the NE.

Application to the Agritrop Use Case. The only specific part concerns the annotation of articles with the AGROVOC thesaurus. Since no gold standard is available, we used the dictionary projection approach with unsupervised entity disambiguation. The integration of disambiguation is still ongoing at the time of writing: Eignethemes must be adapted to compute arbitrary graph embeddings for any standardized SKOS thesaurus, with a technique suited for hierarchies [14].

3.3 Articles Metadata

In addition to text processing steps, the ISSA pipeline requires obtaining the articles' metadata and translating them into RDF. The metadata must contain a URL to download the PDF file of each article, and may contain an identifier, title, authors, date, journal, license, DOI, etc. Depending on the considered archive, metadata may be obtained using various interfaces, commonly a REST API. Therefore, this step will usually require (1) writing a connector to adjust to the archive's API specifics, and (2) adjusting the mapping that lifts the archive-specific metadata to the target RDF model. The ISSA pipeline comes with a connector compatible with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[6] that is largely adopted in scientific data sharing [16].

We have defined an RDF model that represents articles' metadata and content using well-adopted vocabularies: DCMI[7], FRBR-aligned Bibliographic Ontology (FaBiO) [29], Bibliographic Ontology[8], FOAF [18] and Schema.org [19]. A comprehensive description of the RDF representation together with examples are provided in the pipeline's Github repository.[9]

Application to the Agritrop Use Case. In Agritrop, OAI-PMH is used to retrieve the common metadata as well as the abstract and thematic descriptors defined by the curators, that are mapped to RDF using the model described in Sect. 3.1. Given that the text and abstract extracted from the PDF files by Grobid can be of poor quality, we provide a mechanism to coalesce title and abstract retrieved from the metadata with those extracted from full text.

[6] https://www.openarchives.org/pmh/.
[7] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[8] http://bibliontology.com/specification.html.
[9] https://github.com/issa-project/issa-pipeline/blob/main/doc/.

3.4 Integrating All Building Blocks into a Comprehensive Pipeline

Running the Extractors. The pipeline's Github repository provides multiple scripts[10] that orchestrate and automate the processing steps from downloading articles to yielding the resulting RDF KG. To facilitate the deployment, third-party tools Grobid, Annif, Entity-fishing and DBpedia Spotlight are dockerized using official Docker images. In addition, DBpedia Spotlight[11] and Entity-fishing are deployed using pre-trained English and French models.

Generation and Publication of the KG. The translation into RDF of the outputs of each step is carried out using Morph-xR2RML,[12] an implementation of the xR2RML mapping language [27] for MongoDB databases. Thus, the next steps consist of importing the outputs into MongoDB, pre-processing them to filter out unneeded or invalid data, and apply the translation rules with Morph-xR2RML. Lastly, the produced RDF files are loaded into a dockerized Virtuoso OS server deployed using an official Docker image. An additional customizable RDF Turtle file[13] describes the generated RDF dataset using the DCAT [22], VOID [8] and SPARQL-SD [33] vocabularies.

Incremental Updates. After initial publication, periodic invocation of the pipeline can be scheduled to incrementally update the KG with new documents and retrain the Annif models.

Application to the Agritrop Use Case. In the case of Agritrop, the pipeline processed the 12,000 open-access articles in English and French. Annif and the NE extractors were deployed on a virtual machine with 12 CPU cores (2.3 GHz) and 32 GB RAM, the processing took 11 h. MongoDB and Morph-xR2RML were deployed on the same virtual machine. The upload in MongoDB of the documents produced by the NE and descriptor extractors, their pre-processing, the generation of RDF files and their loading into Virtuoso took 1 h 05 m. Additional insights into the dataset generated for Agritrop are given in Sect. 5.

Pipeline Reusability. The pipeline can be customized to meet the needs of any scientific archive and community. The OAI-PMH protocol is very common among scientific archives, such that connecting to archives implementing it should be straightforward. The comprehensive metadata model relies on standard vocabularies and is fully generic. The pipeline is delivered with pre-integrated tools to perform entity-linking against DBpedia, Wikidata, and GeoNames. Yet, new processing steps can easily be defined to leverage other tools and vocabularies suited to specific needs. Finally, the automatic thematic indexing relies on Annif

[10] https://github.com/issa-project/issa-pipeline/tree/main/pipeline.
[11] https://sourceforge.net/projects/dbpedia-spotlight/files/2016-10/en/.
[12] https://github.com/frmichel/morph-xr2rml/.
[13] https://github.com/issa-project/issa-pipeline/blob/main/dataset/dataset.ttl.

that supports numerous models and can be used with arbitrary vocabularies and languages.

4 Visualization and Exploration Services

4.1 Augmented Visualization of Metadata Records

The primary role of an open archive is to provide access to the bibliographic records of the resources it contains. The ISSA prototype meets this need by **enabling users to access an enriched bibliographic view of each open access article in the database**. Beyond merely presenting common article metadata, this service (exemplified in Fig. 2 for the case of Agritrop) visualizes the article abstract where extracted NEs are highlighted and point to the associated knowledge bases (Wikidata, DBpedia, GeoNames, ...). Thematic descriptors automatically extracted with text classification and linked to the considered thesaurus (e.g. AGROVOC) are also shown, along with a cartographic visualization of the places mentioned in the article, linked to GeoNames. Technically, the service consists of a React.js-based web interface and a Node.js server that carries out queries to the semantic index, and is fully generic: adapting the CSS stylesheets suffices to match any other graphical chart.

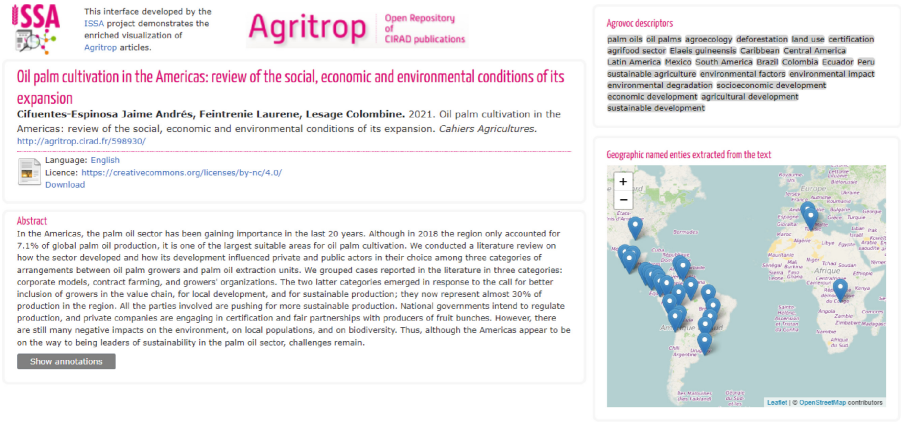

Fig. 2. Augmented visualization of an article's bibliographic records.

4.2 Extraction and Visualization of Association Rules

An association rule is an implication of the form $X \rightarrow Y$, where X is an antecedent itemset and Y is a consequent itemset: transactions containing items in set X tend to contain items in set Y. Each rule is described through its *confidence*, which defines the probability of finding Y in a transaction knowing that X is in the same transaction, and *interestingness*, which defines the serendipity of a rule by penalizing rules with high incidence of antecedent and/or consequent items. Association rule mining is widely used to discover correlations, frequent

patterns, associations or casual structures, and can assist researchers in narrowing down the search for scientific publications.

Using the algorithm proposed in [12], we extract association rules linking the articles' thematic descriptors extracted as described in Sect. 3.1. The mining process casts scientific publications as transactions and thematic descriptors as itemsets. Although the approach helps to reduce and focus the exploration of a dataset, researchers are still confronted with a large set of rules. Therefore, we leverage the potential of visualization to assist the exploration of these rules and thus the discovery of hidden knowledge in the database. In particular, we explore the data using ARViz[14] (Fig. 3), a generic tool designed to support the exploration of association rules via three complementary visualization techniques (i.e. a scatter plot, a chord diagram, and an association graph) providing the distribution of rules over the measures of interest and a focused exploration of (i) **items**, to find and/or describe the rules involving a particular item, and (ii) **rules**, to detect distinguishable association rules that are worth saving for knowledge acquisition.

Application to the Agritrop Use Case. In the analysis, we considered the 3,610 thematic descriptors mentioned in 21,013 articles[15]. To keep only relevant rules, we dropped rules with confidence and interestingness below a given threshold (empirically set to 0.7 and 0.3, respectively), as well as redundant rules (i.e. a rule $A, B, C \rightarrow D$ is redundant if $Conf(A, B \rightarrow D) \geq Conf(A, B, C \rightarrow D)$). The resulting set consists of 20,697 association rules that can be explored using ARViz. Given a antecedent or consequent concept, ARViz dynamically identifies and displays all the relevant associated concepts. For instance, in the current context of the COVID-19 pandemics, researchers might be interested in knowing how strongly the disease relates to other concepts in publications. Thus, we use the association graph view in ARViz to display all the rules involving the concept `COVID-19` (Fig. 3b). The graph provides an intuitive portrayal of antecedent and consequent items involved in the rules (Fig. 3a), where items are represented over on the left and right sides of the screen, and rules are encoded as diamond-shaped nodes placed between the items, which color encodes the measures of interest. This example reveals that `COVID-19` is associated to three consequent concepts: the family `Coronavirinae` of viruses, `pandemics`, and `economic crises`. For the latter, the associated references indeed reveal publications on the resilience of the food sector and agricultural response to the COVID-19 crisis. Concepts co-occurring with `COVID-19` share one or more consequent concepts. This is the case of `food security` that occurs in publications concerned with economic crises and pandemics.

4.3 Exploring Descriptors Co-occurrence

We present below a complementary visualization tool, LDViz[16] [26] which can meet other types of exploration needs and solve complex competency questions.

[14] Accessible at http://dataviz.i3s.unice.fr/arviz/issa.

[15] This includes the 12,000 articles processed by the pipeline, together with articles for which we only have metadata with curator-provided descriptors.

[16] Accessible at http://dataviz.i3s.unice.fr/ldviz.

Use Case 1. The *One Health* initiative [21,23] seeks to unify public, animal and environmental health themes to better understand the development of pandemics and the spread of emerging diseases. In the current context of global climate change, CIRAD researchers wish to figure out publications in the Agritrop open archive that mention both climate change and health (including sub-concepts such as human health, public health, animal health, plant health, etc.), and the time period when these links appeared in CIRAD's research work. To this end, we explore the ISSA semantic index using the LDViz tool which leverages SPARQL queries to explore relevant data through the multiple perspectives delivered by the MGExplorer graphic library. In particular, the tool supports the exploration of relationships within data in cluster and pairwise manners and their distribution over time.

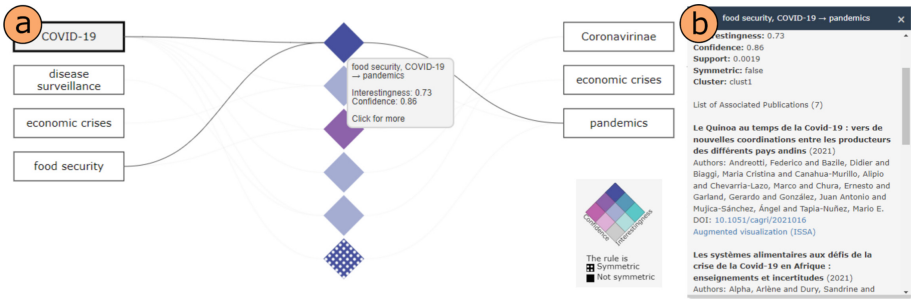

Fig. 3. Visual exploration of (a) association rules involving the COVID-19 concept using ARViz and (b) the publications mentioning the concepts COVID-19, food security and pandemics.

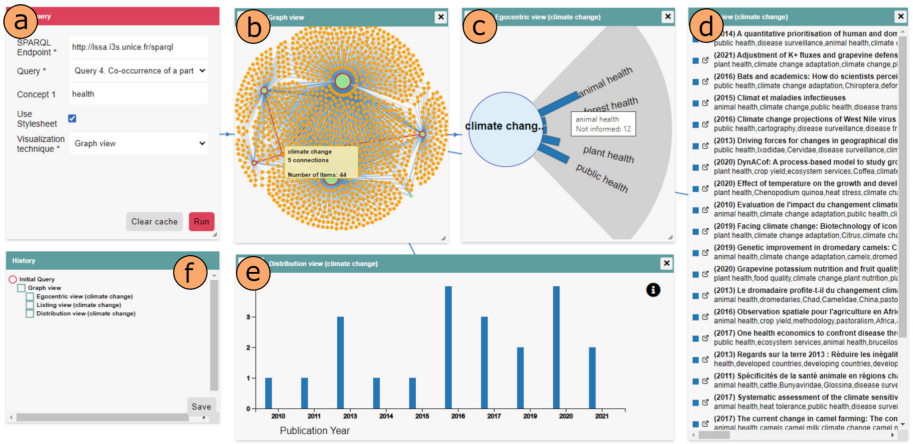

Fig. 4. Visual exploration of health and climate change relationship using LDViz.

To solve the task at hand, we defined a SPARQL query that retrieves the set of articles mentioning `climate change` together with `health` or any narrower or related concept. LDViz proposes a query panel where domain experts can select predefined queries (Fig. 4a) and explore the data through complementary visualization techniques. The exploration starts with a graph view where nodes represent concepts linked together through the scientific publications where they co-occur (Fig. 4b). We continue the exploration with an egocentric view focused on the `climate change` concept since we want to know how it is related to `health`. This shows the different concepts linked to `climate change` and the number of publications where they co-occur. For instance, we can see in Fig. 4c that `climate change` co-occurs mostly with `animal health` in 12 publications. Then, the listing view (Fig. 4d) shows the publications that co-mention `climate change` and `health`, which we can further explore using the other visualizations presented in Sect. 4. Finally, we explore the temporal distribution of those publications (Fig. 4e) where we observe a slightly more intense joint use of those concepts in 2016 and 2020.

Use Case 2. This second use case exemplifies how these tools can be used at institutional and decision-making levels. Public policies are a relevant research subject in CIRAD, as it helps in steering and supporting public decision-making. Thus, we explore the CIRAD publications through the perspective of the `policies` concept to (i) identify the major research areas around public policies, (ii) the ones that are absent or poorly covered, and (iii) the predominant topics across time, which can be contextualized via historical events. We begin the exploration with a graph where green nodes depict the `policies` concept and its narrowers. These are linked to other concepts (in orange) when they co-occur in publications (Fig. 5a). This visualization reveals that CIRAD's major public policy research topic is agricultural policies (central green node). These are strongly linked to `development policies` (Fig. 5d), in line with CIRAD's mandate, as well as to `land policies`. Concepts `water` (Fig. 5c), `food` (Fig. 5b), `forestry` (Fig. 5e) and `environmental policies` (Fig. 5f) are present to a lesser extent while being all related to `agricultural policies`. The time distribution of publications dealing with `environmental policies` reveals a growing interest of research at CIRAD in this field, confirming that their evolution is correlated with relevant world events such as the World Development program (UN) in 2016, the Paris Agreement in 2015, its fifth anniversary in 2020, or the COVID-19 pandemic in 2020.

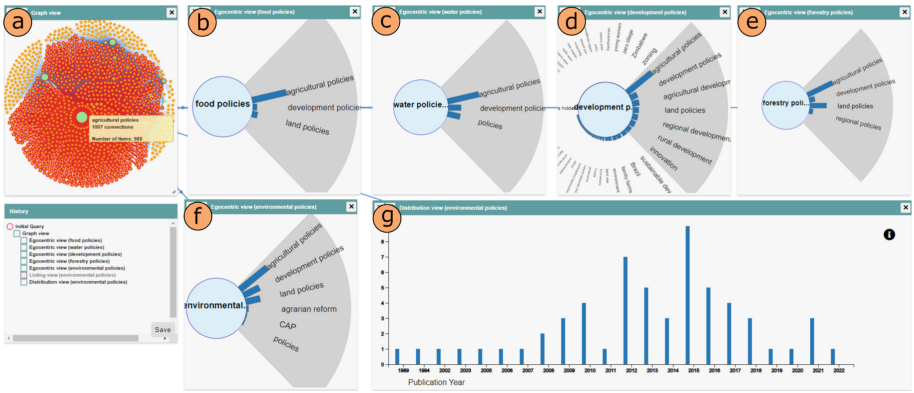

Fig. 5. Visual exploration of scientific publications mentioning any concept of the "Policies" family of descriptors.

5 Source Code, Dataset, Documentation

Source Code Availability. From a technical perspective, ISSA consists of several software components integrated together. Third-party components such as Annif, Grobid, DBpedia Spotlight and Entity-fishing are obtained through their official Docker image distributions on DockerHub.[17] The components developed within the ISSA project are available on Github repositories, licensed under the open-source, free-software Apache 2.0 license, and assigned a DOI that guarantees long-term availability. This information is summarized in Table 1. In particular, the processing pipeline's repository provides multiple scripts that orchestrate and automate the different steps from downloading the articles to running the triple store, together with documentation including deployment instructions, licensing and RDF modelling description.

Sustainability Plan. In the short term, CIRAD wishes to dedicate efforts to the deployment of the ISSA pipeline and visualization tools for production use. This will be the opportunity to assess the quality of the deployment procedure and documentation, and improve them when necessary. Furthermore, a key motivation of the ISSA project is to provide a solution generic enough to be reused with various scientific archives. Therefore, we intend to provide support to communities showing interest in this solution and willing to experiment with it for their own needs. Depending on further funding opportunities, this may range from a best-effort support to more substantial collaborations.

[17] https://hub.docker.com/.

Table 1. Source code developed or adapted for ISSA.

Name	License	DOI	Repository
Processing pipeline	Apache 2.0	10.5281/zenodo.6513983	https://github.com/issa-project/issa-pipeline
Arviz and association rules mining	Apache 2.0	10.5281/zenodo.6511786 10.5281/zenodo.6511146	https://github.com/Wimmics/arviz https://github.com/Wimmics/association-rules-mining
MGExplorer	Apache 2.0	10.5281/zenodo.6511782	https://github.com/Wimmics/ldviz
Article visualization	Apache 2.0	10.5281/zenodo.6510031 10.5281/zenodo.6510029	https://github.com/issa-project/web-visualization https://github.com/issa-project/web-backend

ISSA Agritrop Dataset. The dataset generated by the pipeline for the Agritrop archive is available as a downloadable, DOI-identified RDF dump, and through a Virtuoso OS triple store and SPARQL endpoint. This information is summarized in Table 2 along with basic statistics. The RDF model underlying the dataset is provided in the Github repository.[18] At the time of writing, the URIs are not yet dereferenceable due to on-going security validation procedures required by CIRAD's administrators. In line with best practices [17], the dataset comes with a thorough self-description, comprising (1) licensing, authorship and provenance information, used vocabularies, interlinking and access information, described with Dublin Core Metadata Information, DCAT, VOID and SPARQL-SD.

Table 2. Main facts and statistics about the ISSA Agritrop dataset.

Dataset DOI	10.5281/zenodo.6505847
Downloadable RDF dump	https://doi.org/10.5281/zenodo.6505847
Public SPARQL endpoint	http://issa.i3s.unice.fr/sparql
Documentation	https://github.com/issa-project/issa-pipeline/blob/main/doc/
URIs namespace	http://data-issa.cira.fr/
Dataset URI	http://data-issa.cirad.fr/issa-agritrop
# extracted entities	Named entities: 3.65M, thematic descriptors: 350K
# links to external resources	Wikidata: 2.17M, DBpedia: 1.47M, GeoNames: 152K, AGROVOC: 314K
# RDF triples	66.0M

Dataset Licensing. Being derived from the Agritrop open archive, different licenses apply to the different subsets of the ISSA Agritrop dataset. Articles metadata is provided under the Agritrop open licence[19]. By contrast, article content is ruled by various licenses that consequently also apply to the full text content extracted from the articles and stored in the ISSA dataset. The additional data produced by mining the articles (thematic descriptors, NEs) is published under the Open Data Commons Attribution License 1.0 (ODC-By).[20]

[18] https://github.com/issa-project/issa-pipeline/blob/main/doc.
[19] https://agritrop.cirad.fr/mention_legale.html.
[20] ODC-By license: http://opendatacommons.org/licenses/by/1.0/.

6 Potential Impact and Reusability

Target Audiences and Expected Uses. The ISSA project addresses a widely expressed need in communities that manage open archives, in particular libraries and STI services: provide users with powerful, accurate services to find articles relevant for their goals. The ISSA pipeline not only allows the automatic indexing of articles, but also offers services to find relevant articles by exploiting the richness of their semantic associations. Moreover, adhering to the FAIR principles, the solution can be reused by any community adopting these principles while leaving them free to use terminological references suited to their field. It is therefore aimed at both researchers and specialists in STI, and will be of interest to any person or group in charge of institutional management.

Potential for Reuse. The processing pipeline and visualization services are concrete contributions delivered by the project, designed to be as generic as possible, and successfully tested and deployed in the context of an institutional open archive in production. This technical achievement is a positive indicator of the solution's reusability, and we believe that transferring it to other communities should require only marginal development and adaptation. The adaptation of thematic descriptors extraction may require more substantial work in the absence of a corpus to train supervised models: one should start with an unsupervised model and perform manual validation to bootstrap an annotated corpus of sufficient size for supervised approaches. Furthermore, in line with the dynamics of open science, all developed software is available under an open license, along with all the necessary documentation. Finally, in order to inform, share and transfer our results to other communities, a dissemination workshop was organized in Strasbourg in June 2022 [4].

Impact Assessment. Being the institution that publishes and maintains the Agritrop open archive, CIRAD intends to set up the ISSA pipeline in production as soon as the project will complete (September 2022). This underlines the interest of CIRAD users in the services offered by ISSA, and results from a joint work on application scenarios submitted by CIRAD researchers and scientific information specialists to the ISSA project team. The outcome of this work demonstrates the relevance and flexibility of the prototype for answering competency questions, and the benefit provided compared to traditional search tools integrated into document management platforms. Thus, we are confident that the solution delivered by ISSA can accommodate multiple open archives concerned with similar issues and needs, and help them improve their service offerings.

7 Conclusion and Perspectives

In this article, we have highlighted the challenge of finding relevant publications in the ever-growing body of scientific literature, and presented concrete methods and tools implemented in the ISSA project to deliver services that address this challenge. Leveraging robust, industry-proven tools, we designed a generic,

reusable pipeline for the analysis and processing of articles from an open scientific archive, to produce a semantic index in the form of an RDF knowledge graph. We developed innovative search and visualization services that leverage this semantic index to allow researchers, decision makers or scientific information professionals to explore thematic association rules, co-publication networks, networks of articles with co-occurring topics, etc. We demonstrated the ability of these services to provide answers to real competency questions submitted by researchers.

In the short and middle terms, we plan to continue this work in several ways. First, in terms of data quality evaluation. In particular, evaluating the quality of the text classification models trained with Annif is not trivial. Because of the subjectivity inherent to annotation of documents, the common quality metrics are not so relevant. However we can calculate the similarity metrics between human and machine annotations. Secondly, we intend to apply association rules mining, not only to descriptors, but also to extracted named entities, and assess the quality and usability of these rules. We also wish to enrich our service offering, in particular in terms of bibliometrics and information retrieval, and apply the pipeline to another scientific archive so as to confirm its reusability. Finally, we plan to conduct dissemination activities so that other communities can take up our work and adapt it to their own needs. In the longer term, we believe that the proposed solution could serve as a framework to integrate additional tools and methods, and eventually extract richer, machine-processable knowledge from the mass of human-readable knowledge inherent in scientific archives.

Acknowledgments. This ISSA project is supported by the research infrastructure CollEx-Persée (https://www.collexpersee.eu/projet/issa/).

References

1. Agritrop Portal (2022). https://agritrop.cirad.fr/
2. GO-FAIR Initiative (2022). https://www.go-fair.org/
3. ISSA Project Website (2022). https://issa.cirad.fr/en
4. ISSA Workshop, June 2022 (2022). https://t.co/iYVf7xcdhR
5. OpenMINTED project website (2022). http://openminted.eu/
6. RD Alliance project website (2022). https://www.rd-alliance.org/
7. VisaTM Project Website (2022). https://www.ouvrirlascience.fr/projet-visa-tm/
8. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. W3C Recommendation (2011). http://www.w3.org/TR/2011/NOTE-void-20110303/
9. Arora, A., Garcia-Duran, A., West, R.: Low-rank subspaces for unsupervised entity linking. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 8037–8054. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November 2021. https://aclanthology.org/2021.emnlp-main.634
10. Benedetti, F., Bergamaschi, S., Po, L.: Lodex: a tool for visual querying linked open data, January 2015
11. Budroni, P., Claude-Burgelman, J., Schouppe, M.: Architectures of knowledge: the European open science cloud. ABI Technik **39**(2), 130–141 (2019). https://doi.org/10.1515/abitech-2019-2006

12. Cadorel, L., Tettamanzi, A.G.B.: Mining RDF data of COVID-19 scientific literature for interesting association rules. In: Proceedings of the WI-IAT'20-IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 14–17 December 2020, Melbourne, Australia (2020). https://hal.inria.fr/hal-03084029
13. Caracciolo, C., et al.: The AGROVOC linked dataset. Semant. Web - Interoper. Usabil. Appl. **4**(3), 341–348 (2013). http://content.iospress.com/articles/semantic-web/sw106
14. Chami, I., Wolf, A., Juan, D.C., Sala, F., Ravi, S., Ré, C.: Low-dimensional hyperbolic knowledge graph embeddings. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6901–6914. Association for Computational Linguistics, Online, July 2020. https://aclanthology.org/2020.acl-main.617
15. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems, pp. 121–124 (2013)
16. Devarakonda, R., Palanisamy, G., Green, J.M., Wilson, B.E.: Data sharing and retrieval using OAI-PMH. Earth Sci. Inform. **4**(1) (2010). https://www.osti.gov/biblio/990230
17. Farias Lóscio, B., Burle, C., Calegari, N.: Data on the web best practices. W3C Recommandation (2017). https://www.w3.org/TR/2017/REC-dwbp-20170131/
18. Graves, M., Constabaris, A., Brickley, D.: FOAF: connecting people on the semantic web. Catalog. Classif. Q. **43**(3–4), 191–202 (2007)
19. Guha, R.V., Brickley, D., Macbeth, S.: Schema. Org: Evolution of Structured Data on the Web. Commun. ACM **59**(2), 44–51 (2016). https://doi.org/10.1145/2844544
20. Kettani, F., et al.: Projet VisaTM : l'interconnexion OpenMinTeD - AgroPortal - ISTEX, un exemple de service de Text et Data Mining pour les scientifiques français. In: Ranwez, S. (ed.) IC: Ingénierie des Connaissances, pp. 247–249. Nancy, France, July 2018. https://hal.archives-ouvertes.fr/hal-01839626
21. Lerner, H., Berg, C.: The concept of health in one health and some practical implications for research and education: what is one health? Infect. Ecol. Epidemiol. **5**, 25300 (2015)
22. Maali, F., Erickson, J., Archer, P.: Data catalog vocabulary (DCAT). W3C Recommendation, January 2014. https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
23. Mackenzie, J.: The one health approach-why is it so important? Tropical Med. Infect. Disease **4**, 88 (2019)
24. Menin, A., Cadorel, L., Tettamanzi, A.G.B., Giboin, A., Gandon, F., Winckler, M.: ARViz: interactive visualization of association rules for RDF data exploration. In: Proceedings of the 25th International Conference Information Visualisation (IV), vol. 25, pp. 13–20. Melbourne/Virtual, Australia (2021). https://hal.archives-ouvertes.fr/hal-03292140
25. Menin, A., Cava, R., Dal Sasso Freitas, C.M., Corby, O., Winckler, M.: Towards a visual approach for representing analytical provenance in exploration processes. In: Proceedings of the 25th International Conference Information Visualisation (IV), vol. 25, pp. 21–28. Melbourne/Virtual, Australia (2021). https://hal.archives-ouvertes.fr/hal-03292172

26. Menin, A., Faron Zucker, C., Corby, O., Dal Sasso Freitas, C.M., Gandon, F., Winckler, M.: From linked data querying to visual search: towards a visualization pipeline for LOD exploration. In: WEBIST 2021–17th International Conference on Web Information Systems and Technologies. Proceedings of the 17th International Conference on Web Information Systems and Technologies (WEBIST), Online Streaming, France, October 2021. https://hal.archives-ouvertes.fr/hal-03404572
27. Michel, F., Djimenou, L., Faron-Zucker, C., Montagnat, J.: Translation of relational and non-relational databases into RDF with xR2RML. In: Proceeding of the 11th International Conference on Web Information Systems and Technologies (WebIST), pp. 443–454. Lisbon, Portugal (2015)
28. Michel, F., et al.: Covid-on-the-web: knowledge graph and services to advance COVID-19 research. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 294–310. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_19
29. Peroni, S., Shotton, D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. J. Web Semant. **17**, 33–43 (2012). https://www.sciencedirect.com/science/article/pii/S1570826812000790
30. Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., Varma, M.: Parabel: partitioned label trees for extreme classification with application to dynamic search advertising. In: Proceedings of the 2018 World Wide Web Conference. WWW 2018, pp. 993–1002. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3178876.3185998
31. Science-Miner: entity-fishing (2016–2022). https://github.com/kermitt2/entity-fishing
32. Suominen, O.: Annif: DIY automated subject indexing using multiple algorithms. LIBER Q. **29**(1), 1–25 (2019). https://doi.org/10.18352/lq.10285
33. W3C: Sparql 1.1 service description. W3C Recommendation (2013). https://www.w3.org/TR/2013/REC-sparql11-service-description-20130321/
34. W3C: Web annotation vocabulary. W3C Recommendation (2017). https://www.w3.org/TR/annotation-vocab/
35. Wilkinson, M., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3** (2016)