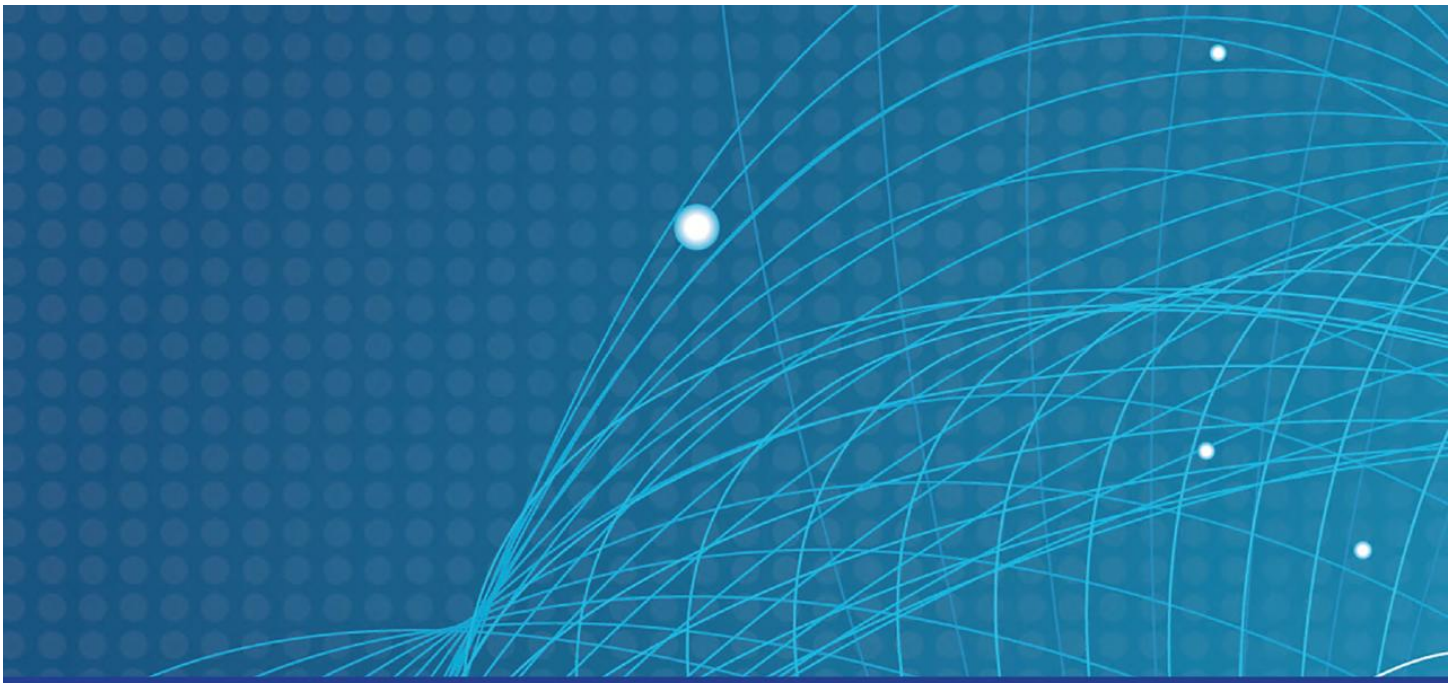


Computational Intelligence in Engineering Problem Solving

SEMANTIC WEB TECHNOLOGIES

RESEARCH AND APPLICATIONS

Edited by
Archana Patel, Narayan C. Debnath,
and Bharat Bhushan

CRC Press
Taylor & Francis Group

Contents

2 Leveraging Semantic Web Technologies for Veracity Assessment of Big Biodiversity Data

Zaenal Akbar, Yulia A. Kartika, Dadan R. Saleh, Hani F. Mustika, Lindung P. Manik, Foni A. Setiawan, and Ika A. Satya

CONTENTS

2.1 INTRODUCTION

In the last couple of years, we have generated and organized an unprecedented amount of data. The number of data created worldwide is huge and grows exponentially. In 2020, we generated about 64.2 Zettabytes, 30 times more than 10 years ago [1]. This phenomenon, known as "big data", where data is characterized by five Vs

DOI: 10.1201/9781003309420-2

(Volume, Velocity, Variety, Veracity, and Value) [2]. Volume refers to the amount of data, Velocity refers to the speed of data generation, Variety refers to different types of data. Veracity and Value refer to the quality of data and benefits of the data respectively. While the first three Vs are used to characterize the key properties of big data, the 4th V (i.e., Veracity) is important to make big data operational [3]. It has become one of the critical factors for creating value because of big data's inherent uncertainty in the form of biases, ambiguities, and inaccuracies [4]. As the consequence, with a large amount of available data, often from a diversity of sources, it is impossible to assess data veracity manually [5]. Methods, algorithms, as well as tools for assessing data veracity automatically, are highly required.

Data veracity refers to the inconsistency and data quality problems, where poor quality data would affect the results of data analyses [3]. The quality of data influences the extraction of useful and valuable knowledge [6], where the presence of uncertainty in the data may negatively impact the effectiveness and the accuracy of the analyses [7]. Therefore, an assessment method for data veracity that deals with uncertain or imprecise data need to look into multiple factors including data inconsistency, incompleteness, ambiguities, as well as deception [8]. More than that, data from multiple sources introduce data conflicts, making data veracity assessment even more challenging [9]. Veracity is also compromised by the occurrence of intentional deceptions such as fake news, malicious rumors, and fabricated reviews [5].

Due to the wide variety of factors and sources that could affect data veracity, we limited our work to the unintentional factor only, specifically those are originated from data consistency and data uncertainty. There are three main sources for data inconsistency, namely the difference in storage format, semantic expression, and value [10]. Storage format refers to the types of the medium where data is stored, most likely in various formats including structured, semi-structured, and unstructured. The semantic expression refers to the way an object is described, where multiple expressions can be used to describe the same object. A value refers to a measured result of the physical quantity, where various types of inconsistency could happen when measuring an object due to human as well as equipment factors. Sources of data uncertainty are also varied, including data collection variance, concept variance, and multi-modality [7]. Variance in data collection could be introduced by environmental conditions as well as data sampling. The same concept might be used not similarly, introducing concept variance. Data complexity and noise from multiple sensors are examples of multi-modality sources for data uncertainty. Dealing with these multitude sources of data consistency and uncertainty is about filtering out or amending the data through data cleaning and data reduction [11], a necessary step toward successfully big data analytics.

2.1.1 Motivation and Research Challenges

Biodiversity research is organized into domains that cover distinct spheres of biodiversity knowledge, e.g., taxonomy, geographical distribution, or functional traits of organisms [12]. In this work, we focus on the studies of the distribution

of life across space and time, providing a key link between organisms and their environment, also known as biogeography. To study the link, a biodiversity information system would hold various information about the organisms and their observed environments. Typically, biodiversity data contains observations of the occurrences of specific species that can be identified by a specific taxonomic name at a specific geographic location at a specific time [13]. As an example, the Global Biodiversity Information Facility (GBIF)[1] has recorded 1 billion records of species occurrences in 2018.[2] The record is constantly growing as data are provided by more than 1,600 institutions across the globe. Another example is Pl@ntNet,[3] a citizen science project that helps users to identify plants based on pictures provided by citizens. The project has published more than 6,6 million records of occurrences in two datasets.[4] Another example is eBird,[5] a citizen science project that enables volunteered observers to report bird observations. The project has published an observation dataset with more than 700 million records of occurrences.[6]

In general, biodiversity data can be produced in two methods, data collection and data mobilization [14]. In the first method, data generation will be started with a field survey where researchers would visit (predetermined) locations to observe species occurrences or even to collect specimens. In this step, information that is related to the locations, as well as the observation time, will be also recorded. In the case of citizen science projects, observation can also be performed by amateurs, normally by using instruments such as wireless devices or portable microscopes [15,16]. For specimen collection, the specimens will be preserved in a special room such as a laboratory. If not done yet, each record or specimen will be identified further to determine the correct taxonomic name. Then, all those data will be entered into a biodiversity information system. In the second method, data will be extracted from preserved specimens or works of literature such as checklists and taxonomic monographs.

Our work is motivated by the relatively complex procedures to produce high-quality biodiversity data as described above. Based on multiple factors involved during data production, we hypothesized that biodiversity data tends to be noisy. The noise which will affect data veracity could come from multiple factors as follow:

1. First, location factors. At least, two locations are involved in data generation, namely field observation, and laboratory identification. This situation contributes to data corruption such as missing values. For example, information about the habitat of a species that supposed to be collected from the field was forgotten. In the case of extraction of distributional information from preserved specimens, many important characteristics, e.g., plant growth form, vegetative height, or stem specific density could be missing [14].
2. Second, time factors. the id time differences between field observation to lab identification. Process errors such as data redundancy are highly possible.

3. Third, human factors. error also possible to have happened in the last phase, data entry. While data entry is performed independently it will be performed manually by humans so data entry errors can happen.

We believe that these three factors have a tremendous impact on the data veracity of biodiversity data. To the best of our knowledge, there is no existing work yet to investigate this problem.

2.1.2 Research Objectives

In this chapter, we propose an automatic method to assess the veracity of biodiversity data through data consistency analysis. Data consistency analysis can be performed from a variety of perspectives, including database development, computing strategy, and data science [10]. In this work, we use the perspective of data science especially big data management, where data that are scattered in distributed sources is required to be integrated, where data consistency across multiple sources is important to ensure high quality integrated data. The proposed method utilizes a data mapping solution to align data from multiple sources into a predefined structure by using a standardized vocabulary in a way data consistency can be measured and compared. Data mapping solutions bring many advantages, for example, exposing underneath schema of relational databases [17]. Exposing the schema of multiple databases is important in big data analytics. The expose would help users to have a better understanding of the structure of the databases, and at the same time help users to optimize queries to those databases. Data mapping solutions also would enable dataset trustworthiness by exposing the provenance of mapping quality [18]. It is more effective to assess and refine data mapping definitions than to assess and refine the quality of a dataset directly. Furthermore, a data mapping definition can be refined further to improve the quality of data [19]. Based on the assessment of data mapping quality, if a problem (for example data types inconsistency) is detected then a mapping definition refinement can be suggested to automatically improve the mappings.

Our research question is as follows: "How have data inconsistency in structures, value types, and granularities affected the veracity of open biodiversity data?". As reflected in this research question, we would like to investigate the impact of three sources of data inconsistency on the veracity of biodiversity data. First, data structure consistency refers to how elements of data are structured. Second, value types consistency refers to how a similar element data is used across multiple sources. Third, data granularity consistency refers to how specific a similar element data has been described. The rest of the chapter will be organized as follows: relevant and related works including our contributions will be explained in Section 2.2. Our method to measure data consistency will be described in Section 2.3. After presenting our results in Section 2.4, we summarize our chapter in Section 2.5.

2.2 RELATED WORK

We align our work with two prominent research areas, namely big data veracity analysis and data quality analysis of biodiversity data. In this section, we describe several related works from each area and outline our contributions.

Even though the veracity dimension of big data remains under-explored compared to the other dimensions, many works have been done to analyze it in several big data applications. Reference [4] proposes a big data veracity index by defining three main theoretical dimensions of veracity, namely objectivity, truthfulness, and credibility. The index was used to assess systematic variations of textual information across multiple datasets. They found that each dimension might contribute to the overall quality to a different degree, and therefore should be assigned different weights. Since the multi-modality of information sources could amplify the veracity of data, reference [9] proposes cross-modal truth discovery by predicting the truth label of claims through linkage analysis of various events from multiple sources. The approach was able to infer the reliability of sources with no or little prior knowledge. In another work, reference [20] proposes a platform to estimate data veracity by extracting entities, relations, and structures of claims to be combined in a way the veracity label of data and trustworthiness scores of the sources can be determined. Multiple methods can be used to determine the veracity data of electronic medical records, including process mining and using ontology [8]. In the process mining method, data quality will be assessed by mapping the chronological time/date within the records. Ontology can be used to share quality metrics. Standardized terminology can also provide data correction for misspelt words in unstructured text fields. And most recently, the big data should be transformed into smart data where data must be appropriately sorted, structured, and analyzed [6,11]. Smart data aim to filter out or amend imperfect data through data cleaning and data reduction, for example for dealing with data redundancy or contradiction.

Data quality is also becoming a major issue in the field of biodiversity science. Numerous factors could affect data veracity, including observation error, expertise, and reliability of the primary data collectors, possible data corruption during secondary data management and analysis, and any other factor that might increase uncertainty [21]. The integration of biodiversity data deals with the availability, quality, and interoperability of data which are mostly based on disaggregated data types [14]. One of the challenges is missing or inconsistent data items that can be solved with the data imputation method where a value will be estimated (using logical and statistical approaches) to replace the missing data. The increase of "big unstructured data" has highlighted the discrepancy in global data availability between data quantity and data quality [22]. It is necessary to do benchmarking big unstructured data against high-quality structured datasets, as well as developing purpose-specific rankings to assess data quality. A controlled vocabulary and data annotation could improve data quality and fitness

for use [23]. Also, the communities which have the necessary expertise to validate, curate, and improve data from diverse sources should be integrated into the data [24]. This approach would enable researchers to engage effectively and efficiently with vast volumes of complex data, to contribute through simple curatorial actions to improve digital knowledge. Furthermore, integrating and transforming biodiversity data into a knowledge graph requires extensive data cleaning and cross-linking [13]. For instance, converting data from multiple sources into a specific format requires multiple steps. Even though a set of declarative mapping rules can be used to align entities from multiple sources to a targeted scheme [25], it remains challenging to reconcile entities across multiple sources.

In line with these two broad research areas explained above, we outline our main contribution as a method to assess data veracity in biodiversity data, such as sources, comparison of veracity types. Based on the definition of sources of noises from [6], our work assesses the "attribute noises" which can be explained as a corruption of data in the values of the input attributes. In this type of veracity source, the factors can be erroneous attribute values, missing values, or incomplete attributes. In contrast with existing works that mainly rely on machine learning approaches, we lay our work on the fundamental approach for data integration, namely data mapping. We map element data from multiple sources into one defined data structure in a way the noise will prevail. Our work is different from the data mapping approach in [8], which was utilized for the correction of misspelt words only. Our method goes beyond that, namely assessment of data consistency in structure, data values, as well as data granularity.

2.3 METHOD

In this section, we introduce our method to assess data veracity of open biodiversity data. We measure and compare data inconsistency across multiple data sources. Three sources of data inconsistency will be investigated, namely data structure inconsistency, data value types inconsistency, and data granularity inconsistency. First, we define our data definitions as an approach to representing data from multiple sources into one generic schema. Second, we describe how to measure three types of inconsistency from the obtained mappings. And finally, our research procedures will be explained at the end of this section.

2.3.1 Data Definitions

To be able to measure the three types of data consistency, we introduce several definitions and formalizations as follows:

1. **Vocabulary**: Vocabulary is a collection of data attributes that can be used to describe an object. Each attribute has a name and an expected type of value. As an illustration, to describe a biological specimen, it is necessary to have a few data attributes such as the name of the specimen, where and when the specimen was collected, and so on. Further, an

attribute "name" should have a textual value, an attribute "date" should have a date value, etc. In the field of biodiversity, several existing vocabularies have been used widely. One of them is Darwin Core,[7] a data standard for publishing and integrating biodiversity information [26]. We use this vocabulary due to its wide adoption.

2. **Dataset**: Dataset is a collection of data objects, where each object is described with one or more attributes that are available in the selected vocabulary.
3. **Dataset vocabulary**: Dataset vocabulary is a collection of data attributes, where each object is described with one or more attributes that are available in the selected vocabulary. It is worthy to mention that every dataset could use less or a greater number of attributes available in the vocabulary.

2.3.2 Data Consistency Analysis

After introducing the basic definitions of our data mapping solution, we constructed our data consistency measurement as follows:

1. **Data structure consistency**: The first measurement applies to the dataset level, meaning that it can be used to measure consistency across datasets. Two datasets will be called consistent if every related data objects in both datasets utilize similar data attributes. One dataset may have a richer structure than the other.
2. **Data type consistency**: The second measurement is applicable to attribute level, meaning that it will be used to compare attributes across data objects within a dataset or to compare related data objects across datasets. Two data objects will be called consistent if both objects utilize a similar data type for their relevant attributes.
3. **Data granularity consistency**: The third measurement is applicable at the data value level, meaning that it will be used to compare values of related attributes of data objects within a dataset or across multiple datasets. We use a semantic similarity[8] metric to measure the distance between two values. Semantic distance is a metric to measure how far a concept is from other concepts in a knowledge organization system. By identifying concepts in a data value of an attribute, we can map each concept to a knowledge organization system such that the semantic distance between them can be measured. If an attribute is data granularity consistent then all values of this attribute should be mapped to the same concept.

We model the value of an attribute by using a Simple Knowledge Organization System (SKOS),[9] a vocabulary and data model for expressing knowledge organization systems for data referencing and reusing. For example, the concept "location" will be modeled in SKOS as shown in Table 2.1. There are three important semantic relations, namely "broader", "narrower", and "related". The relation "broader" and

TABLE 2.1
Modelling of Concept "Location" Using SKOS

Subject	Predicate	Object
ex:Island	rdf:type	skos:Concept
	skos:prefLabel	"Pulau"@id
	skos:prefLabel	"Island"@en
	skos:related	ex:StateProvince
	skos:related	ex:Country
ex:StateProvince	rdf:type	skos:Concept
	skos:prefLabel	"Provinsi"@id
	skos:prefLabel	"State Province"@en
	skos:altLabel	"Province"@en
	skos:broader	ex:Country
ex:Country	rdf:type	skos:Concept
	skos:prefLabel	"Negara"@id
	skos:prefLabel	"Country"@en
	skos:narrower	ex:StateProvinc

"narrower" are transitive relations to represent if a concept is broader or narrower than others respectively. The relation "related" is a reflexive relation to represent that a concept is related to the other and vice versa. The model can also be visualized in a graph representation as shown in Figure 2.1. This graph depicts that "StateProvince" has a broader concept so-called "Country" and "Country" has a narrower concept so-called "StateProvince". An "Island" is related to both "StateProvince" and "Country".

2.3.3 Data Mapping Procedures

After describing our formalization to measure data consistency in the previous sub-sections, now we introduce our data mapping procedures. Figure 2.2 shows our mapping procedures that consist of three main activities as explained in the following sub-sections.

1. **Data crawling and extraction**: In this first activity, data will be crawled and extracted from several sources. Since most of the data are available over the Web, a web-scraping method will be employed to extract the required data. The input of this activity is a list of Uniform Resource Locator (URL). The output will be a collection of files, where each file contains data as a tuple in the form of (key, [values]). In each tuple, a "key" has a list of zero, one, or more "values".
2. **Data mapping**: From every tuple obtained in the previous activity, its "key" will be mapped to a relevant attribute of the selected vocabulary. The relevancy will be determined by users that have a wide variety of

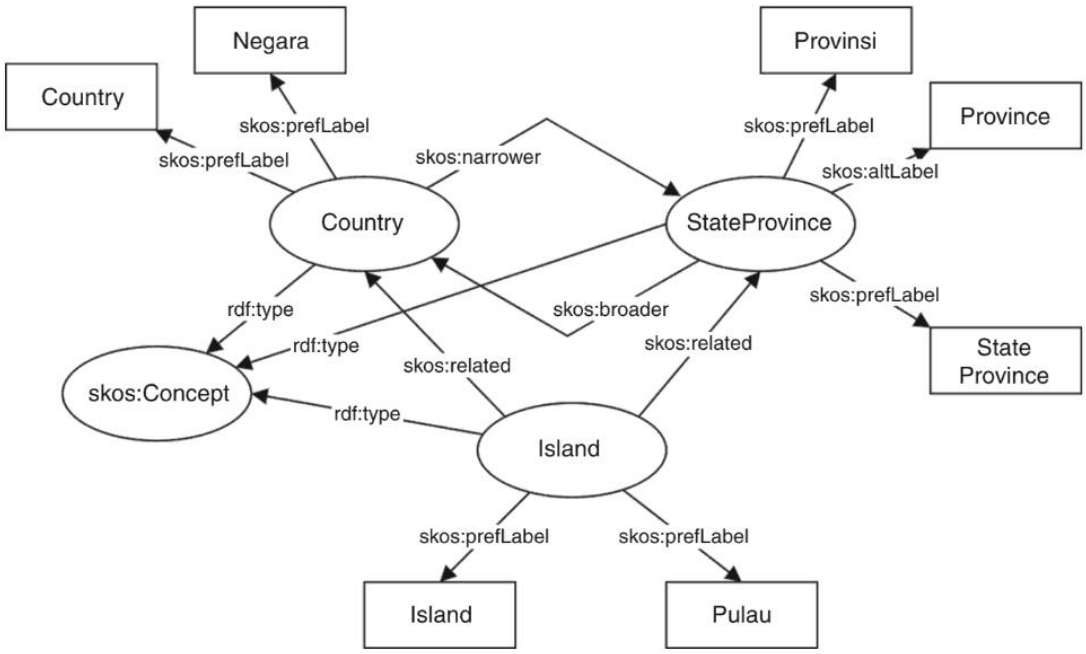


FIGURE 2.1 Visualization of concepts "location" using SKOS.

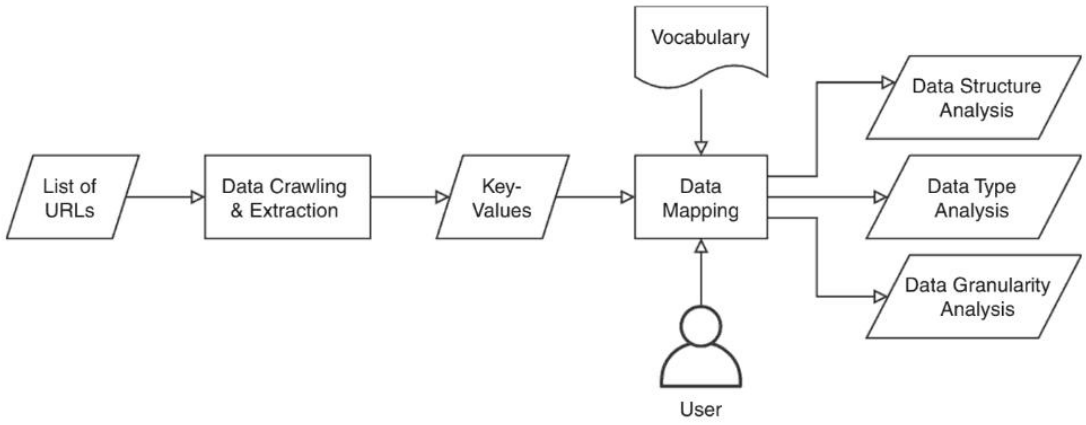


FIGURE 2.2 Research methodology.

expertise in data science. As output, a list of (key, attribute) will be produced. In case of no relevant attribute can be identified, then a tuple (key, Null) will be produced.

3. **Data analysis**: For analysis, we constructed a third collection of tuples based on the defined mapping in the previous activity. Technically, the process will be performed as (key, [values]) + (key, attribute) = (attribute, [values]). To answer our research question, three types of analyses will be performed, explained as follow:
   a. To measure data structure consistency, the tuples (key, attribute) will be compared from one dataset to another. We expected to be able to identify which attributes are widely used and which ones are not.

b. To measure data type consistency, the data type of "values" of selected "key" in tuples (key, [values]) will be matched with the data type of "values" defined for "attribute" in the relevant tuples (attribute, [values]).
c. To measure data granularity consistency, "values" of selected "attribute" in tuples (attribute, [values]) will be aligned with concepts from a defined knowledge system. The distance between one concept to another will be computed to measure the granularity of the relevant "attribute" which is associated with "key".

Apart from the above three types of analyses, we also would like to collect tuples (key, Null) to be analyzed further for vocabulary enrichment in the future.

2.4 RESULT

In this section, we describe our results, discuss our findings, and summarize the lessons learned from the findings.

2.4.1 Dataset

To assess open biodiversity data, we collected data from multiple sources with a few limitations as follows:

1. We limited our biodiversity data that are related to data observation of species occurrences at a specific location at a specific time.
2. We limited our data collection to the publicly available data, published on the Web.
3. From each dataset, we are limited by the publicly available fields only. To the best of our knowledge, not all fields of the database are opened to the public.

The summary of the dataset is shown in Table 2.2. We collected more than 60,000 records of species occurrences from nine sources. It covers botanical as well as zoological data.

2.4.2 Dataset Vocabulary

As explained in Section 2.3, we need to construct a dataset vocabulary based on the mapped "attributes" of the selected vocabulary i.e. Darwin Core Terms. As result, there are 72 terms were used in our collections of datasets as shown in Figure 2.3. The top three terms belong to class "Taxon", "Location", and "Event" respectively.

2.4.3 Data Structure Analysis

To analyze data structure across multiple datasets, we computed how each dataset uses our vocabulary and which attributes are not used in each dataset as shown

TABLE 2.2
Collection of Datasets

No.	URL	# Specimen
1	http://ibis.biologi.lipi.go.id/	10,629
2	http://ibis.biologi.lipi.go.id/mzb/	1,749
3	http://bankbiji.krbogor.lipi.go.id/katalog	114
4	http://sindata.krcibodas.lipi.go.id/	1,969
5	http://ipbiotics.apps.cs.ipb.ac.id/	1,074
	http://indobiosys.org/	14,222
6	http://ipt.biologi.lipi.go.id/	17,250
7	Herbarium of Andalas University	1,128
8	Museum Zoologi Bogor	14,043
9	Tambora Muda Indonesia	
	Total	62,178

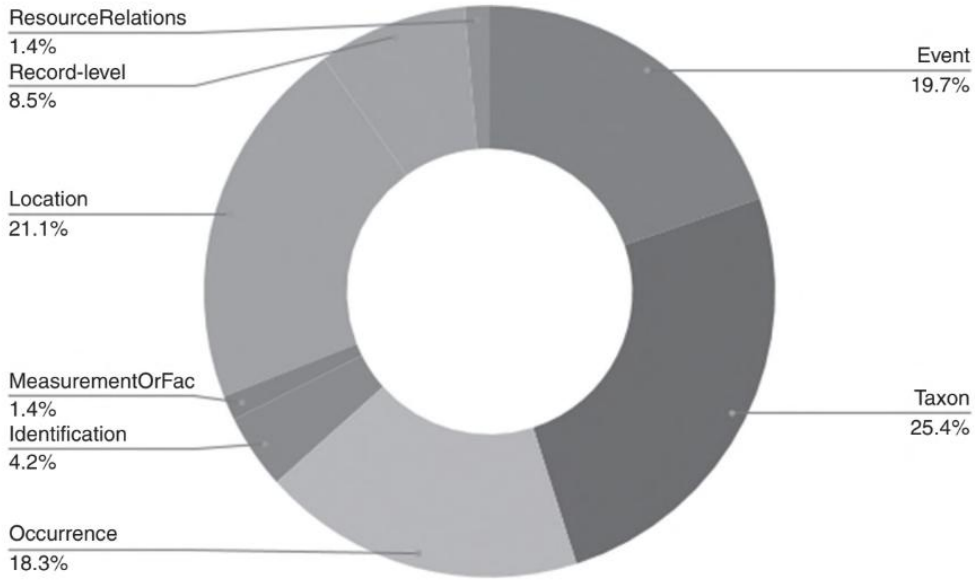


FIGURE 2.3 The proportion of 72 Darwin Core terms that are used as our dataset vocabulary.

in Figures 2.4 and 2.5 respectively. As shown in Figure 2.4, class "Location" was widely used across datasets but has an uneven distribution. Class "Taxon" is also widely used across datasets with better distribution. Other classes were used only partially, for example, class "Event" was used in eight datasets only, class "MeasurementOrFact" was used in two datasets, or class "ResourceRelationship" was used in one dataset only. When identifying which terms are used in each dataset, we obtained results as shown in Figure 2.5. The ratio of terms used that are available in our vocabulary across multiple datasets was very low. Only two datasets were utilized above 50%, two datasets use less than 40%, 1 less than 30%, 2 less than 20%, and even less than 10% for one dataset.

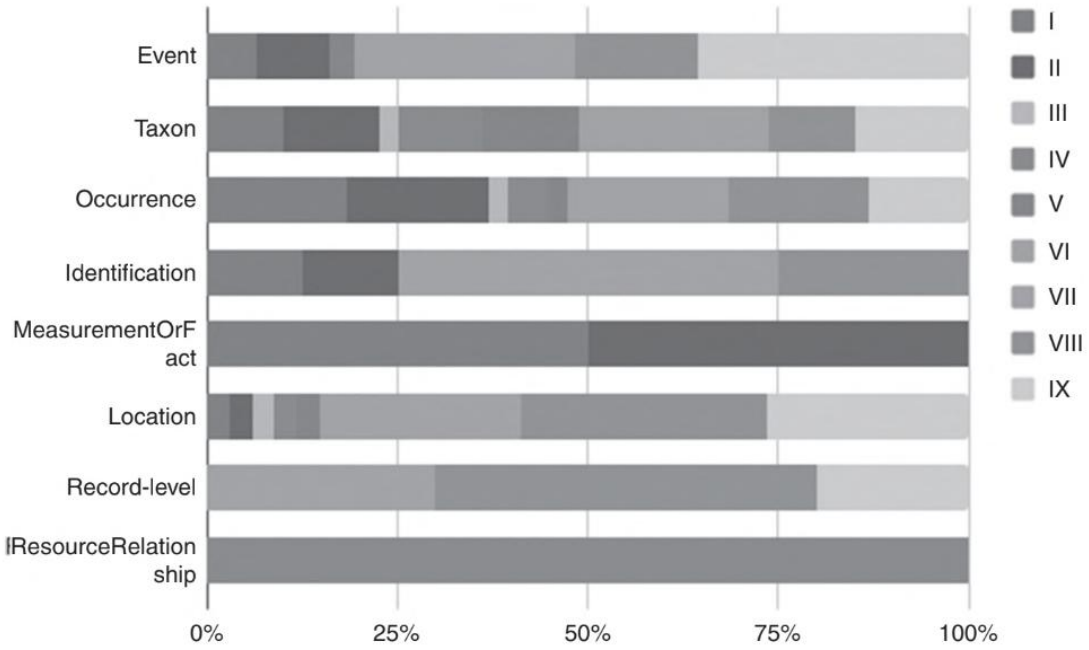


FIGURE 2.4 Proportion of our dataset vocabulary across datasets.

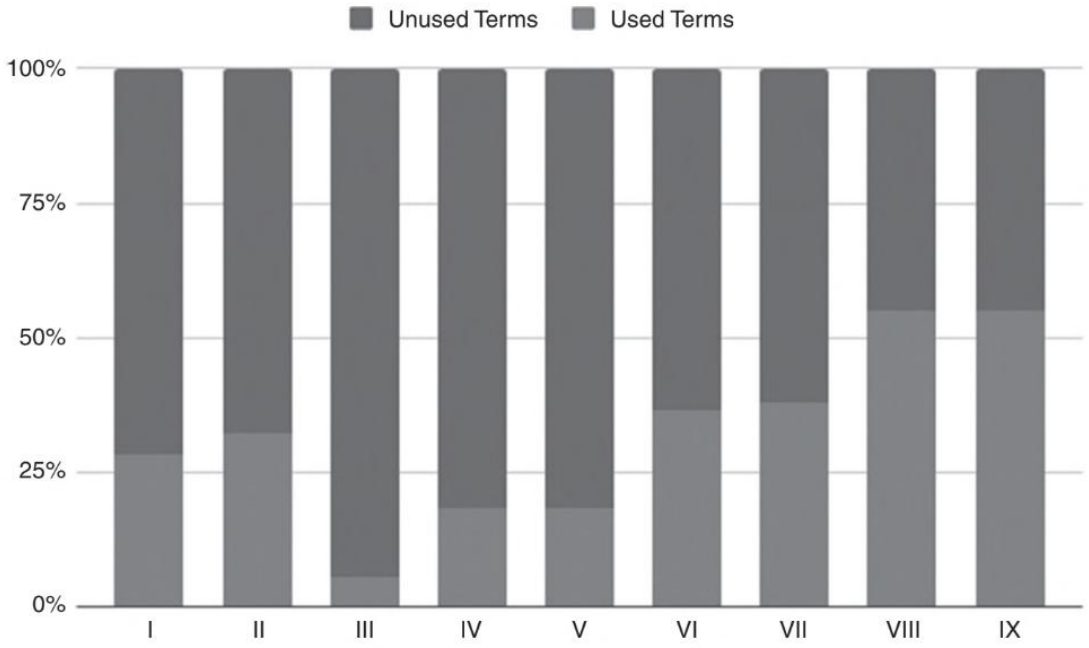


FIGURE 2.5 Proportion of used and un-used attributes across datasets.

From these two results, we explain the situation as follows:

1. Data structure if very heterogeneous
2. High consistency is found in several classes, namely "Taxon" and "Event" due to their relatively equal distribution across multiple datasets. This indicates that the potential for data integration can be started with these two classes.

Even though we found a relatively low consistency across multiple datasets, we would like to outline potential biases during our data collection and processing as follow:

1. data collection bias refers to how data were collected by different organizations according to the research question that they would like to answer. For example, one organization is probably concerned only about specific species in specific locations and therefore terms that are related to the habitat would not be recorded.
2. data mapping bias refers to a possible error caused by an incorrect alignment between "key" of data to "attribute" in the selected vocabulary.

2.4.4 Data Type Analysis

To analyze data type consistency across multiple datasets, we selected the attribute "eventDate" of class "Event" which according to the definition of Darwin Core can be used to specify the date-time or interval during which an "Event" occurred. It is recommended to use a date that conforms to ISO 8601-1:2019.[10] The standard provides an unambiguous representation of dates and times to avoid misinterpretation of numeric representations of dates and times across different conventions. The attribute was used in seven datasets, but in two datasets the value of the attribute was presented as a combination with values from other fields, and therefore they were discarded. We also disregarded records that have empty values in their corresponding attributes. The result from five datasets is shown in Figure 2.6.

As shown in Figure 2.6, seven formats were employed, where one format represents date intervals, one represents a specific date with time intervals and the other represents a specific date and time. One dataset uses four different formats, while four others use only one format. We also found that only one format was

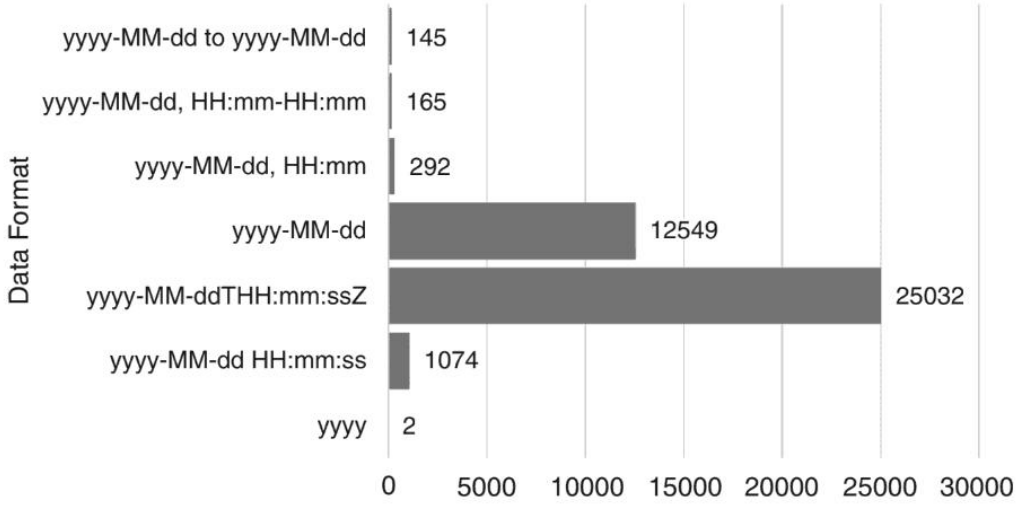


FIGURE 2.6 Number of records that use different data formats in our datasets.

used consistently across two datasets, namely (yyyy-MM-ddTHH:mm:ssZ). Only one format specified a time zone designator (Z) explicitly, which is turn out to be employed by most of the data records.

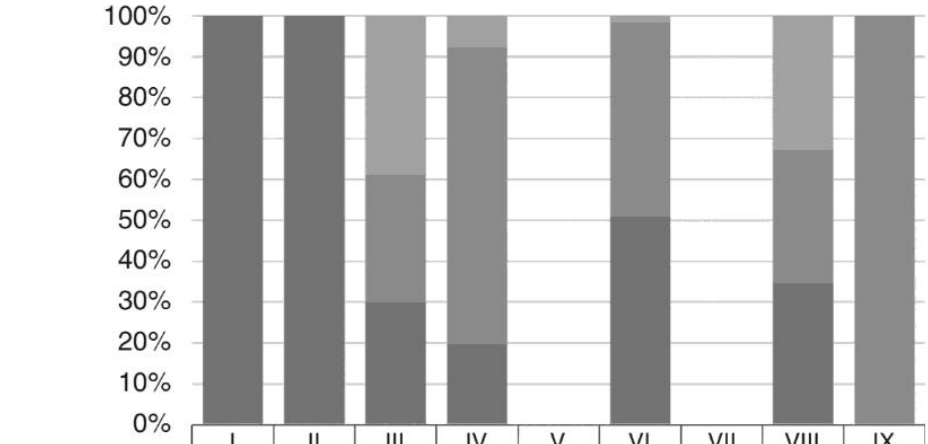
We explain the situation as follows:

1. Multiple formats can be used across multiple datasets that still conform to a standard (such as ISO 8601).
2. A time zone designator is not adopted wide enough, which could lead to multiple issues when integrating data from multiple sources.

2.4.5 Data Granularity Analysis

To assess data granularity consistency, we selected three attributes in our dataset vocabulary that refer to a geographical location. The attributes are "country", "locality", and "island" from class "Location". An attribute "country" refers to the name of the country or major administrative unit in which a "Location" occurs, "locality" refers to the specific description of the "Location", and "island" refers to the name of the island on or near which the "Location" occurs. These three attributes were selected because, based on our mapping, they were used across multiple datasets. Three datasets were used attribute "country", five and one datasets were used attributes "locality" and "island" respectively. From every selected attribute, we constructed a knowledge system as depicted in Figure 2.1. The knowledge system consists of three concepts, namely "Country", "StateProvince", and "Island". After that, we analyze how those concepts were utilized in our datasets and the result is shown in Figure 2.7.

As shown in Figure 2.7, the three concepts to represent a geographical location were used differently across multiple datasets. Concepts of "Country" and



	I	II	III	IV	V	VI	VII	VIII	IX
Island	0	0	38,835	7,3913		1,291		32,671	0
StateProvince	0	0	31,068	72,754		47,673		32,419	100
Country	100	100	30,097	19,855		51,036		34,91	0

FIGURE 2.7 Portions of datasets that use different concepts to specify a geographical location.

"StateProvince" were used consistently in two and one datasets, respectively. Two datasets used a different concept, namely "GeoCoordinate". In the remaining four datasets, those three concepts have been used simultaneously. We describe this situation as follows:

1. The level of granularity is varying across multiple sources. Forcing to follow a specific granularity level is hard, especially in the case of data mobilization. In this case, the information related to a geographical location can be missing due to various factors. For example, the name of a place has been changed.
2. The use of attribute "GeoCoordinate" is becoming popular and should be used as the first option. In the case of data mobilization, it is important to have a global mapping from textual name to this attribute such that the granularity level can be consistent.

2.5 CONCLUSION

In this chapter, we introduced a method to assess data veracity of open biodiversity data. The method performs a consistency analysis on three important sources of data consistency, namely data structure, data type, and data granularity. To the best of our knowledge, our work is the first one that investigated these sources thoroughly. The analysis was performed in the context of the big data ecosystem, where data is distributed across multiple sources, presented in various formats, and maintained by multiple organizations. Our main objective was to assess the veracity of biodiversity data automatically which can be used further to improve the quality of data.

Three sources of data consistency were investigated. First, data structure analysis, applied at the datasets level, was intended to measure how a defined vocabulary was utilized across multiple datasets. Second, data type analysis, applied at the data attribute level, was intended to measure how the data type of an attribute is used within a dataset or across multiple datasets. Third, data granularity analysis, applied at the data value level, was intended to measure how multiple concepts were used in values of selected attributes. As our datasets, we collected publicly available biodiversity data more than 60,000 records of species occurrences that are available in nine distributed data sources. Our analysis was conducted in several systematic steps, namely:

1. Data collection, where biodiversity data from multiple sources were collected. In most cases, web-scraping techniques were used to extract the relevant pair of (key, [values]) for every element data that was presented on a website.
2. Data mapping, where every extracted "key" is mapped to the most suitable attribute in a selected vocabulary to produce a pair of (key, attribute). We used the Darwin Core as our vocabulary due to its wide adoption in the biodiversity area.

3. Based on the mapping, we constructed a final collection of (attribute, [values]). After that, several statistics regarding data consistency were computed.

As a result, we obtained:

1. There is a high number of terms (72 in total) that were utilized across nine datasets. The terms are classified in Taxon, Location, Occurrence, Event, Record-level, Identification, MeasurementOrFact, and ResourceRelations.
2. The terms were utilized imbalance across sources. A class of terms such as Occurrence is widely used but three sources underused it compared to the others. The same case with terms in class Location, widely used but under usage by five sources.
3. There is a high diversity of ways to represent a specimen. Even though we obtained 72 terms in our vocabulary, only two data sources utilized more than 50%.
4. We also identified data inconsistency in data type and data granularity. We identified different ways to define data value for a term with type DateTime. Different granularities were used to specify the value for a location.

In conclusion, due to the high inconsistency found in all three sources, performing analysis of big biodiversity data requires more effort in the step of preprocessing step. Data integration as the first step toward analytics requires the implementation of recent technologies to fill the gap. Those technologies are including:

1. Data value completion, which is including link prediction by using machine learning.
2. Data modelling can be used to suppress the level of inconsistency that can be developed further to extend the existing vocabulary. For example, to have a better representation of location or habitat.

ACKNOWLEDGEMENT

This research was supported by the Research Center for Informatics, National Research and Innovation Agency, Indonesia. We would like to thank all the members of the Knowledge Engineering Research Group for their valuable suggestions and feedback.

REFERENCES

1. Statista.com, "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025," 2021. https://www.statista.com/statistics/871513/worldwide-data-created/.
2. M. Younas, "Research challenges of big data," *Serv. Oriented Comput. Appl.*, vol. 13, pp. 105–107, 2019, doi:10.1007/s11761-019-00265-x.
3. B. Saha and D. Srivastava, "Data quality: The other face of big data," in *2014 IEEE 30th International Conference on Data Engineering*, Mar. 2014, pp. 1294–1297, doi:10.1109/ICDE.2014.6816764.

4. T. Lukoianova and V. L. Rubin, "Veracity roadmap: Is big data objective, truthful and credible?" *Adv. Classif. Res. Online*, vol. 24, no. 1, p. 4, Jan. 2014, doi:10.7152/acro.v24i1.14671.
5. M. García Lozano et al., "Veracity assessment of online data," *Decis. Support Syst.*, vol. 129, p. 113132, Feb. 2020, doi:10.1016/j.dss.2019.113132.
6. D. García-Gil, J. Luengo, S. García, and F. Herrera, "Enabling smart data: Noise filtering in big data classification," *Inf. Sci. (Ny).*, vol. 479, pp. 135–152, Apr. 2019, doi:10.1016/j.ins.2018.12.002.
7. R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: Survey, opportunities, and challenges," *J. Big Data*, vol. 6, no. 1, p. 44, Dec. 2019, doi:10.1186/s40537-019-0206-3.
8. A. P. Reimer and E. A. Madigan, "Veracity in big data: How good is good enough," *Health Informatics J.*, vol. 25, no. 4, pp. 1290–1298, Dec. 2019, doi:10.1177/1460458217744369.
9. L. Berti-Equille and M. L. Ba, "Veracity of big data: Challenges of cross-modal truth discovery," *J. Data Inf. Qual.*, vol. 7, no. 3, pp. 1–3, Sep. 2016, doi:10.1145/2935753.
10. P. Shi, Y. Cui, K. Xu, M. Zhang, and L. Ding, "Data consistency theory and case study for scientific big data," *Information*, vol. 10, no. 4, p. 137, Apr. 2019, doi:10.3390/info10040137.
11. J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, *Big Data Preprocessing: Enabling Smart Data*, Cham: Springer International Publishing, 2020, doi:10.1007/978-3-030-39105-8.
12. J. Hortal, F. de Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle, "Seven shortfalls that beset large-scale knowledge of biodiversity," *Annu. Rev. Ecol. Evol. Syst.*, vol. 46, no. 1, pp. 523–549, Dec. 2015, doi:10.1146/annurev-ecolsys-112414-054400.
13. R. D. M. Page, "Ozymandias: A biodiversity knowledge graph," *PeerJ*, vol. 7, p. e6739, Apr. 2019, doi:10.7717/peerj.6739.
14. C. König, P. Weigelt, J. Schrader, A. Taylor, J. Kattge, and H. Kreft, "Biodiversity data integration—The significance of data resolution and domain," *PLoS Biol.*, vol. 17, no. 3, 2019, doi:10.1371/journal.pbio.3000183.
15. S. Kelling, "eBird: A human/computer learning network to improve biodiversity conservation and research," in *Twenty-Fourth IAAI Conference*, Jul. 2012, p. 11.
16. D. R. Hardison et al., "HABscope: A tool for use by citizen scientists to facilitate early warning of respiratory irritation caused by toxic blooms of Karenia brevis," *PLoS One*, vol. 14, no. 6, p. e0218489, Jun. 2019, doi:10.1371/journal.pone.0218489.
17. M. Rodriguez-Muro and M. Rezk, "Efficient SPARQL-to-SQL with R2RML mappings," *J. Web Semant.*, vol. 33, pp. 141–169, 2015.
18. T. De Nies, A. Dimou, R. Verborgh, E. Mannens, and R. Van de Walle, "Enabling dataset trustworthiness by exposing the provenance of mapping quality assessment and refinement," in *The 4th International Workshop}on Methods for Establishing Trust of (Open) Data (METHOD2015)*, 2015, p. 7.
19. A. Dimou et al., "Assessing and Refining Mappingsto RDF to Improve Dataset Quality," in *The Semantic Web - ISWC 2015*, vol. 9367, M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. D'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, and S. Staab, Eds. Cham: Springer International Publishing, 2015, pp. 133–149.
20. M. L. Ba, L. Berti-Equille, K. Shah, and H. M. Hammady, "VERA: A platform for veracity estimation over web data," in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 2016, pp. 159–162, doi:10.1145/2872518.2890536.

21. S. S. Farley, A. Dawson, S. J. Goring, and J. W. Williams, "Situating ecology as a big-data science: current advances, challenges, and solutions," *Bioscience*, vol. 68, no. 8, pp. 563–576, Aug. 2018, doi:10.1093/biosci/biy068.
22. E. Bayraktarov et al., "Do big unstructured biodiversity data mean more knowledge?" *Front. Ecol. Evol.*, vol. 6, p. 239, Jan. 2019, doi:10.3389/fevo.2018.00239.
23. T. Robertson et al., "The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet," *PLoS One*, vol. 9, no. 8, p. e102623, Aug. 2014, doi:10.1371/journal.pone.0102623.
24. D. Hobern et al., "Connecting data and expertise: A new alliance for biodiversity knowledge," *Biodivers. Data J.*, vol. 7, p. e33679, Mar. 2019, doi:10.3897/BDJ.7.e33679.
25. Z. Akbar, Y. A. Kartika, D. Ridwan Saleh, H. F. Mustika, and L. Parningotan Manik, "On using declarative generation rules to deliver linked biodiversity data," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, 2020, pp. 267–272, doi:10.1109/ICRAMET51080.2020.9298573.
26. J. Wieczorek et al., "Darwin Core: An evolving community-developed biodiversity data standard," *PLoS One*, vol. 7, no. 1, p. e29715, Jan. 2012, doi:10.1371/journal.pone.0029715.

NOTES

1 https://www.gbif.org/, Accessed: 30/10/2021.
2 https://www.gbif.org/news/5BesWzmwqQ4U84suqWyOQy/, Accessed: 30/10/2021.
3 https://plantnet.org/, Accessed: 30/10/2021.
4 https://www.gbif.org/publisher/da86174a-a605-43a4-a5e8-53d484152cd3, Accessed: 30/10/2021.
5 https://ebird.org, Accessed: 30/10/2021.
6 https://www.gbif.org/dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e, Accessed: 30/10/2021.
7 https://dwc.tdwg.org/, Accessed: 30/10/2021.
8 https://en.wikipedia.org/wiki/Semantic_similarity, Accessed: 30/10/2021.
9 https://www.w3.org/2004/02/skos/, Accessed: 30/10/2021.
10 https://en.wikipedia.org/wiki/ISO_8601, Accessed: 30/10/2021.

8 Semantic Annotation of Objects of Interest in Digitized Herbarium Specimens for Fine-Grained Object Classification

Zaenal Akbar, Wita Wardani, Taufik Mahendra, Yulia A. Kartika, Ariani Indrawati, Tutie Djarwaningsih, Lindung P. Manik, and Aris Yaman

CONTENTS

DOI: 10.1201/9781003309420-8

8.1 INTRODUCTION

Machine learning techniques, especially supervised ones, have been widely used in numerous applications, including natural language processing, biological image classification, stock market analysis, self-driving cars, and precision agriculture [1]. Furthermore, with the increased availability of data and computational resources, techniques in many applications have been widely accepted [2]. In biodiversity research, machine learning techniques have been used for multiple purposes, such as plant sciences, mainly on images of herbarium specimens [3]. Plant morphology, growth and development, the ecological interactions of plants with herbivores, and other related cases may now be analyzed quickly, precisely, and easily using machine learning. These techniques could also be used to assess the global change biology by processing herbarium data known to have biases over space, time, and phylogeny [4]. Machine learning can also help the basis of biological research, such as improving the accuracy of species identification [5]. The discovery of phenological patterns on unprecedented scales has also been made possible by machine learning and combined data from herbarium specimens and spatiotemporal data retrieved from specimen labels [6].

When it comes to herbarium specimens, it is widely known that this is the most valuable data source for biodiversity research. From about 3,100 herbaria around the World, there are a total of 390 million botanical specimens.[1] Extinct, uncommon, endemic, and common botanical specimens are all preserved in herbarium collections to serve as a reference for future study. Many efforts have been performed to digitize the specimens and share them, so they can be used by researchers all over the world. An example of the portals that provide digitized specimen data is the Integrated Digitized Biocollections (iDigBio) Portal, which has a collection of approximately 131 million digitized specimens.[2]

Without a doubt, these data sources offer a great potential to help us learn more about biodiversity, and machine learning techniques are one approach to achieve that. Multiple works have been done in this area. For example, to identify leaves and other components from digitized herbarium specimens [7], and to discover patterns and trends of plant-herbivores interactions [8–11]. Further biological analysis can also be performed, for example, to determine the driver of shifting interactions such as phenological change, distributional shifts, and urbanization [10]. Another example is using machine learning to segment plant tissues in herbarium specimen images and removes the background pixels [12].

For supervised training of machine learning models, large training datasets are necessary [13]. Data variability is an essential aspect of machine learning, especially for object identification or object classification tasks. Therefore real-world

image variation is essential to ensure the results [14]. But on the other hand, using uncontrolled natural images might be seriously misleading, dangerously leading in the incorrect direction. Labeled data is made up of a large set of representative photos that have been labeled or highlighted with the relevant features. An extensive and accurate labeled dataset, the ground truth, is required for training the algorithm [15]. Recent advancements in machine learning approaches, such as deep learning techniques, can generate features automatically, which saves feature engineering costs, but in return, may require larger volumes of labeled data [16]. The size and quality of training datasets will affect the quality of the trained models [17]. Due to those requirements, it is now essential to share the raw data and, most importantly, the annotated/labeled data.

8.1.1 Annotation of Digitized Herbarium Specimen

Figure 8.1 shows two images of annotated digitized herbarium specimens. Each image consists of at least two types of information:

1. The sheet on the right bottom side of each image depicts the specimen's label, which includes information about the spatiotemporal dimension of the specimen, as well as when and where the specimen was collected. It also contains information about the person who has collected it and taxonomic data on the specimen.
2. The images themselves depict parts of a plant such as leaves, branches, flowers, and fruits. The yellow rectangles alongside the annotations indicate where the images have been annotated. The first image has six annotations, while the second image has 21 annotations scattered around their leaves. As can be seen, the annotated areas' size is not uniform, and two annotations can overlap the others.

FIGURE 8.1 Two examples of annotated digitized herbarium specimens.

In this work, we focus on the second type of information, namely parts of the specimen that has been annotated. The annotations in this example will be used to identify objects of interest within images of herbarium specimens. Furthermore, using machine learning techniques, the annotations will be utilized to build a model for automatic object classification.

8.1.2 Motivation

Dealing with such a vast amount of data requires a systematic approach. Machine learning techniques are sometimes unfamiliar to scientists who work with data. Having tools that assist scientists in processing data and revealing its potential is critical. For example, plant scientists may use an object detection application programming interface (API) to assist an object detection pipeline in detecting morphological features of a plant specimen [18]. A workflow for generating high-quality image masks for segmentation tasks can also be made available to scientists outside of the domain to help them [12]. Another example is providing a tool to annotate an image with specific pre-defined labels [11].

Despite the efforts to make digitized herbarium specimens more accessible, there are still several issues and challenges that remain. One challenge is finding efficient pre-processing techniques to produce a learning system that can deal with data collected from various sources. This problem arises as a result of data being scattered over multiple areas, systems, and applications. The "meaning" of the data may differ from one source to another, which may significantly impact the quality of the machine learning outcomes [2]. The discrepancies label in the annotated images is one example. In order to perform correctly, machine learning requires a sufficient amount of data training with the proper label. There are several challenges introduced by labeling data that have motivated our work, as follows:

1. Various annotation data formats. There are a number of picture annotation tools available, such as LabelMe[3] [19] and VGG Image Annotator (VIA)[4] [20], each with its own data format. There are a couple of widely used data formats, such as the JSON-based Common Object in Context (COCO) data format[5] and the XML-based Visual Object Classes (VOC).[6]
2. Labels tend to be noisy. Label noise can significantly impact the performance of deep learning models [21]. Since erroneous predictions might influence decision, so labeling requires domain expertise with a wide range of knowledge and the capacity to make precise label.
3. A label is applied individually without considering the relations to another annotation. As an illustration, in a digital image of a herbarium specimen, as can be seen in Figure 8.1, we can annotate the part of the plants like leaves and the damage caused by herbivores on the leaves. Both annotations are self-contained for each case and do not consider the relation between them. However, the annotation for the damages indicates that the damages happened on the leaves. Such that we can infer the relations that the damages are parts of the leaves.

The rest of the paper will be organized as follows: Section 8.2 lists and discusses a few related works and outlines our contribution. After that, our research methodology will be explained in Section 8.3. Finally, Section 8.4 presents our results and discusses our findings before concluding our work and explaining a few future works in Section 8.5.

8.2 RELATED WORK

In this section, we describe a few related works from various topics. Then, we align our work to two broad research topics: ontology and semantic annotation for biodiversity research. Finally, at the end of this section, we outline our contributions.

8.2.1 Ontology for Biodiversity Research

Biodiversity science, like most other fields, has been flooded with a huge amount of data. Biological specimens that have been collected in various herbaria across the world have become a valuable data source for biodiversity research. Furthermore, technology has also introduced a new data source, so-called born-digital [22], in which data is collected digitally without collecting a specimen first. Combining these digital data with traditional data sources like data from in situ and remote sensors, community data resources, biodiversity databases, and data from citizen science have pushed this field into Big Data Era [23]. Besides the Volume of the available data, this research area also deals with the Variety and Veracity of the data. The two latter mentioned dimensions are related to and determine the quality of the data. Data collected by multiple organizations and stored in various formats are two examples of how semantic technologies such as ontology, could play a key role in big data analytics.

Ontologies play a crucial role in improving data aggregation and integration across the biodiversity domain in this area. They can be used to describe physical samples, sampling processes, and biodiversity observations that involve no physical sampling [24]. It has been predicted that the data will become less centralized, but the need for cross-species queries will become more common [25]. That is why ontologies would help scientists to achieve that. For example, Plant Ontology (PO) has been widely used to describe plant anatomy and morphology, as well as stages of plant development [26]. A simplified version of PO also can be used to drive a question answering dialog between non-expert users and a knowledge-base about Capsicum [27]. Another widely used ontology is the Darwin Core (DC), a standard for sharing data about the occurrence of life on earth and its associations with the environment [28]. It provides terminology for describing multiple types of information from an organism, such as taxonomic, location, and sampling protocol. It can be used to not only record the occurrence of a species at a specific time and location but also to manage alien species [29] and as a hub to connect data across multiple biodiversity information systems [30].

8.2.2 Semantic Annotation for Biodiversity Research

Data annotation provides multiple advantages. First, it allows for data enrichment by embedding more information. Resulting in more efficient data discovery, so the data can be found, accessed, integrated, and re-used. Digital images of specimens from natural history collections must be bound to semantically rich data across museums to provide a unified user experience [31]. Semantic enrichment of herbarium specimen digital data would reduce the orthographic data variances, particularly for person and place names [32]. Further, data standardization and harmonization can be accomplished simply by using dictionary mapping to annotate data set columns [33]. It will improve discovery, interoperability, re-use, traceability, and reproducibility of the data.

Data annotation also can be used to encode knowledge into data by labeling objects of interest from texts, digital images, or videos as an example. The encoded labels can be used to enhance the data as well as serve as training material for automatic data extraction and classification. For example, multiple annotations of named entities in historic biological texts have been used to fine-tune a machine-learned classifier [34]. In that case, an annotation framework was developed based on a modified version of the Model-Annotate-Model-Annotate (MAMA) cycle. As a result, links between the exploration of biodiversity literature and document retrieval can be provided. The utilization of optical character recognition (OCR) would automate the acquisition process of herbarium specimen metadata [35]. Combined with other specimen image analysis services, the approach would provide a high degree of automation for information extraction from herbarium specimens. An approach for semi-automated extraction of named entities from natural history archival collections also can be developed by using the semantic annotation of the collections [36]. In that case, digital images of the field book will be annotated by drawing a bounding box over the image and attaching additional information. A tool for marine image annotation would increase efficiency and effectiveness in the manual annotation process [37]. In another case, images can be annotated by selecting features of interest (e.g., flowers, birds, or patterns) as tokens with bounding boxes from the images [38]. Finally, tokens can be associated with properties or traits (e.g., colors, behaviors) derived from pre-defined domain ontologies.

8.2.3 Contributions

In line with the two research areas discussed above, we outline our contributions as follows:

1. In contrast with existing works that primarily introduced new annotation tools, our solution would utilize existing annotation tools. We focus more on the annotation format produced by each annotation tool and align them to a unified schema to achieve a common annotation. To the best of our knowledge, none of the existing works have looked into this situation.

2. Similar to other existing approaches, our solution would also allow annotating digitized images based on the entities and properties of pre-defined ontology. But, instead of treating the annotations as independent entities, our solution would also consider the relationship between them. These inferred relations can be used to fine-tune the performed classification tasks.

8.3 METHOD

This section describes our research methodology, followed by our method to develop a uniform schema, and how to align annotations produced by various tools to the schema through mapping rules.

8.3.1 Methodology

Figure 8.2 shows our research methodology. It consists of three main activities as follows:

1. Schema development activity. It is a method for identifying and formalizing information taken from digitized herbarium specimens, as well as its structure. Interviewing domain experts, in this case botanists, is how the activity is carried out. As a result, a schema will be created that covers plant characteristics.
2. Labeling activity. It is a process of marking all regions of interest and their relevant labels for each herbarium specimen image. Domain experts can perform this task using any existing tools they usually use. This activity produces two types of information, are regions of interest (typically by using bounding boxes) and their relevant labels.
3. Mapping activity. It is a process to align marked objects of interest with a pre-defined schema. As a result, at the end of the process, a unified annotation will be obtained.

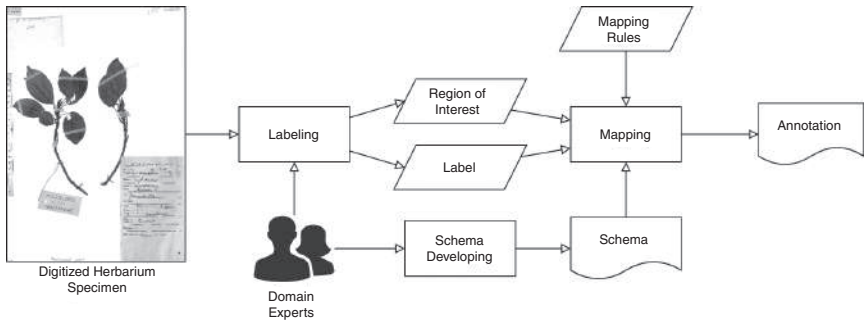


FIGURE 8.2 Research methodology.

It is important to note that the produced labels will be highly task-dependent. It will depend on the objective of the classification tasks at hand. For example, when the classification task considers only damages caused by herbivory, there is no need to label the other cause of damages. However, if we also consider the spatiality of the damages, it will be necessary to label parts of the plant, such as leaves where damages are found.

8.3.2 Schema Development

An annotation schema will be created to represent annotations uniformly across numerous annotation tools and to enrich the annotation by defining relationships between entities within an annotated object. The majority of the concepts are derived from Plant Ontology (PO), a widely used ontology for describing plant anatomy, morphology, and developmental stages [26].

8.3.2.1 Entities

As the representation of objects or things in the domain of interest, we identify several types of entities that can be extracted from digitized herbarium specimens.

8.3.2.1.1 Plant Morphological Entities

Plant morphology refers to the physical appearance of a plant. Physical characteristics of a plant that can be found in a herbarium specimen include leaves, fruits, flowers, and so on. Leaves, in particular, are an essential feature for species delimitation and recognition [18]. Further, trait information about area, perimeter, shapes, colors, textures of leaves can provide important insight into plant species' ecology and evolutionary history [7]. The morphology of plants, such as leaves, flowers, fruits, bark, and branches is ideal for image-based plant species identification [5]. The challenges lie in the diversity of similar features. For example, leaf morphologies of plants native to specific regions will be different from the plant from other regions [11]. Therefore, it is necessary to have a standardized way to share information about these features to ensure they can be consumed and appropriately re-used.

The PO adopt a Gene Ontology (GO) data model to cover flowering plants in general. The PO is ideal for sharing knowledge among scientists who know the issue but is not always understandable by non-expert users [27]. It has been widely used as a common reference ontology for plant structures and development stages. In the same vein, we created a modest but powerful ontology that met our requirements. We started with a simple ontology to improve Capsicum species literacy.[7] The ontology is a small subset of the PO.

8.3.2.1.2 Plant-Animal Interactions Entities

Plant-animal interactions can be seen as a process that has immediate and delayed effects on both entities that interact [39]. An interaction entails an enormous diversity of outcomes depending on interaction type (predation, symbiosis, parasitism, mutualism, commensalism) [40]. Herbarium specimens contain additional information, like nutrients, defense compounds, herbivore damage, disease lesions, and

sign of physiological processes that capture ecological and evolutionary responses [4]. In particular, herbivory damage interactions data can be utilized to uncover various global change drivers across a diversity of insect herbivore-plant associations.

Examples of plant-animal interactions, in this case the damage caused by herbivory, are shown in Figure 8.1. As we can see, several damages on leaves can be identified visually. Based on these interactions, further analysis can be performed, for example, to identify different types of interactions and can be associated with different types of animals that contribute to the interactions [11].

8.3.2.2 Entity Relationships

Besides super-class and sub-class relationships, we would like to outline several critical other types of relationships.

Table 8.1 shows three essential relationships that can be used to represent how entities within digitized herbarium specimens are related to each other. While the first relation relates to morphological relationships, the last two relationships relate to spatial proximity between entities.

8.3.3 Mapping Rules

Declarative mapping, and more precisely mapping rules, establish relationships between various schemas of multiple data sources and a common schema. The relationships align data elements from each source to a single common target, as well as the appropriate structural and data type transformations. Declarative mappings are available in a variety of formats, including the widely used and language-independent tables and spreadsheets [41] to highly structured formats such as R2RML[8] and RML [42]. RML[9] defines mapping rules from heterogeneous data structures and serializations other than relational databases, for example, CSV, XML JSON, to the RDF dataset.

8.4 RESULT

In this section, we list and discuss our results. First, we describe the characteristics of our dataset, followed by the description of our schema. After that, the mapping procedure will be explained before discussing our findings.

TABLE 8.1
Entity Relationship

No.	Relation	Meaning
1	part_of	This relation represents if an entity is part of another entity. It is a core relation to describe a part and its whole.
2	adjacent_to	This relation represents if an entity is in contact with or in spatial proximity to another entity.
3	located_in	This relation represents if the location of an entity is within the location of another entity.

8.4.1 Dataset

To construct our dataset, we digitized the herbarium collection from the Herbarium of Bogoriense.[10] The digitization began with the identification of a collection of specimens suspected to interact with insects, represented by various damages on the specimen. The dataset can be explained as follow:

1. We digitized herbarium specimens from *Excoecaria agallocha*, which belongs to the genus *Excoecaria of Euphorbiaceae*. In total, we obtained 244 specimen sheets.
2. From each digitized sheet, we asked experts to annotate them with three types of damages:
   a. pre-processing (damages that occur before the specimen were collected)
   b. during-process (damages that occur during specimen collection and or during drying and mounting on a sheet)
   c. insects (damages that occur due to insects during preservation)
3. The VGG Image Annotator was used to conduct the annotation. An annotator would draw a bounding box to indicate damage and choose one possible source of the damage.
4. It is important to note that the number of damages on each sheet can be multiple and of different types, overlapping damages in a sheet are also possible.

Table 8.2 shows the size of our multi-labeled dataset. Each sheet of herbarium specimen in our dataset can have several labels. In most cases, all labels occur. Our dataset has around three thousand labels in total, which are divided into three categories. The distribution of label count for each type of label is shown in Figure 8.3. The first three figures depict the distribution of labels for every type of damage, pre-processing, during processing, and insects, respectively. As we can see, most of the damages are up to six, mainly for every type of damage. Finally, the last image depicts the distribution of damages, and as we can see, the average damages are 6, 4, and 3 for every type of damage, respectively.

TABLE 8.2
Collection of Datasets

No.	Types of Labels	# Object of Interest
1	Damages during pre-processing	1,420
2	Damages during process	1,069
3	Damages caused by insects	882
	Total	3,371

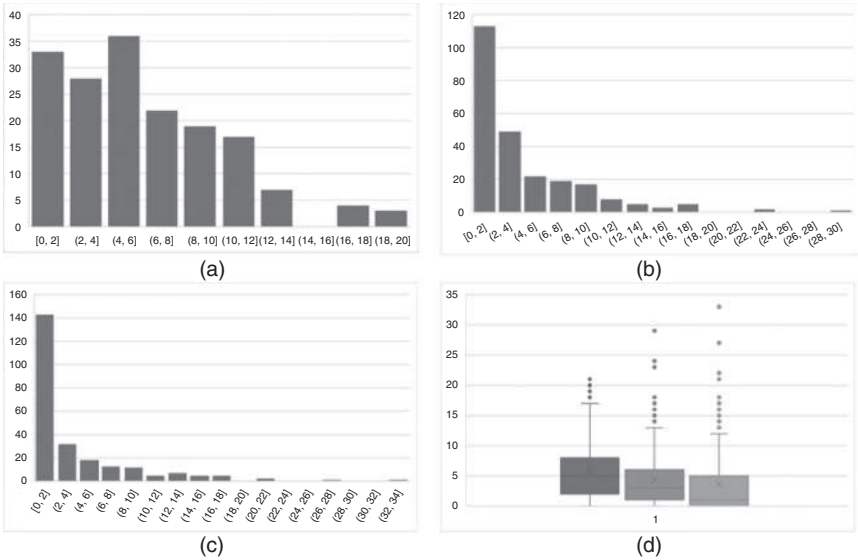


FIGURE 8.3 Label distribution based on damages (a) pre-processing (b) processing (c) insect (d) distribution of all three labels.

8.4.2 Schema

We developed our schema by re-using terms from multiple existing ontologies as follows:

1. The plant structures, we adopted the schema from the OntoCapsicum,[11] an ontology to improve species literacy of Capsicum, which covers morphological characteristics of seeded plants in general [27].
2. The animals, since our research object is plants, we focused on herbivores, animals whose primary food source is plant-based. Herbivores can be classified further into frugivores (fruit eaters), granivores (seed eaters), nectivores (nectar feeders), and folivores (leaf eaters).
3. The interaction refers to the interaction between herbivores and plants which can be characterized by defense mechanism mark or damage on the specimens.
4. The interaction mark, the specimen's spatial dimension can be viewed as a mark of interaction (based on the image perspective).
5. The temporal dimension of the interaction represents the time when an interaction took place.

Figure 8.4 depicts our schema for annotating herbarium specimens, modeled using the Protégé Editor [43]. The core object in a herbarium specimen is the "Plant, " which will be described by the plant's main morphological traits, such as "Flower, " "Leaf, " "Fruit, " and "Stem". The animal that interacts

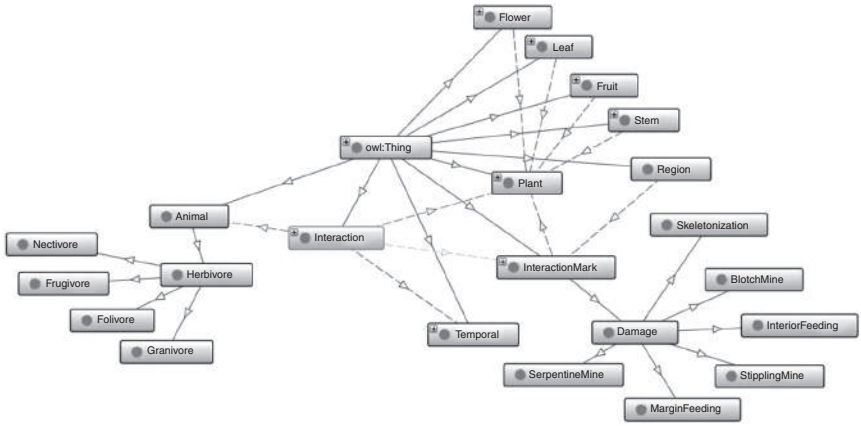


FIGURE 8.4 Schema for annotating herbarium specimens.

with the plant is then represented, in this case is a "Herbivore" entity, which can be further classified as a "Nectivore, " a "Frugivore, " a "Folivore, " and a "Granivore". Further, an entity "Interaction" will represent the interaction between a herbivore and a plant as reflected in a herbarium specimen. Next, "InteractionMark" can be used to identify the interaction on the specimen, which can be "Damages" or "DefenceMechanism". Multiple types of damages can be identified, such as "MarginFeeding", "InteriorFeeding", "Skeletonization", "BlotchMine" and so on. To reflect its spatial location inside the specimen image, the marks will be represented as "Region.". An interaction can also be annotated further with "Temporal" to represent when the interaction happened. The schema currently has 44 entities, 6 object attributes, and 14 data properties in its initial version.

8.4.3 Data Mapping

As mentioned earlier in this section, we would like to annotate the damages found on herbarium specimens and classify the damages based on the time when the damage occurred. Three classes were defined: prior-processing, during-processing, and after-processing (damages caused by insects).

```
<#InteractionMapping> a rr:TriplesMap;
  rml:logicalSource [
  rml:source "Batch-1-Updated.json";
  rml:referenceFormulation ql:JSONPath;
  rml:iterator "$._via_img_metadata.[*]"
];
  rr:subjectMap [
  rr:template "http://lipi.go.id/herbarium/{filename}";
  rr:class hso:Interaction;
];
```

```
  rr:predicateObjectMap [
  rr:predicate hso:hasRegion;
  rr:objectMap [ rr:parentTriplesMap
<#InteractionMarkMapping>;
  rr:joinCondition [ rr:child "filename"; rr:parent
"filename"; ];
];
].
<#InteractionMarkMapping> a rr:TriplesMap;
  rml:logicalSource [
  rml:source "Batch-1-Updated.json";
  rml:referenceFormulation ql:JSONPath;
  rml:iterator "$._via_img_metadata.[*]"
];
  rr:subjectMap [
  rr:template "http://lipi.go.id/herbarium/
{filename}-{size}";
  rr:class hso:InteractionMark;
];
  rr:predicateObjectMap [
  rr:predicate hso:hasRegion;
  rr:objectMap [ rr:parentTriplesMap <#RegionMapping>;
  rr:joinCondition [ rr:child "filename"; rr:parent "shape_
attributes.id"; ];
];
].
```

Figure 8.5 shows a snapshot of our mapping rules using RML in combination with the Function Ontology (FnO).[12] We use JSON files generated by the VGG Image Annotator for the input files. We ran into a few issues when generating mapping for the files, mainly because the RML has limited support for nested data, such as nested objects in a JSON object [44]. We found it challenging to map objects in an array because no specific field can distinguish between members of the array and map them to their parent object. To solve this issue, we pre-processed the input files by inheriting the identification field from the parent object into members of the array in the child object. In this way, the parent object can be linked to each array member. As shown in Figure 8.5, several mappings are defined as a collection of "TriplesMap". A "logicalSource," a "subjectMap," and one or more "predicateObjectMap" are all defined in each definition. The relationships between entities were defined using "parentTriplesMap" from "objectMap" of the source to other related definitions.

The RMLMapper[13] is used to generate the annotation based on the updated input files. Figure 8.6 shows a snapshot of the annotation produced by the mapper, and Table 8.3 list the number of corresponding statements/triples. Out of 244 annotated digital herbarium specimens, we generated 21,058 triples using the current mapping rules. The majority of the triples are linked to the spatial information of damages detected on specimen images.

```
<#InteractionMapping> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "Batch-1-Updated.json";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$._via_img_metadata.[*]"
  ];
  rr:subjectMap [
    rr:template "http://lipi.go.id/herbarium/{filename}";
    rr:class hso:Interaction;
  ];
  rr:predicateObjectMap [
    rr:predicate hso:hasRegion;
    rr:objectMap [ rr:parentTriplesMap <#InteractionMarkMapping>;
      rr:joinCondition [ rr:child "filename"; rr:parent
"filename"; ];
    ];
  ].

<#InteractionMarkMapping> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "Batch-1-Updated.json";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$._via_img_metadata.[*]"
  ];

  rr:subjectMap [
    rr:template "http://lipi.go.id/herbarium/{filename}-{size}";
    rr:class hso:InteractionMark;
  ];

  rr:predicateObjectMap [
    rr:predicate hso:hasRegion;
    rr:objectMap [ rr:parentTriplesMap <#RegionMapping>;
      rr:joinCondition [ rr:child "filename"; rr:parent
"shape_attributes.id"; ];
    ];
  ].
```

FIGURE 8.5 A snapshot of our mapping rules.

```
@prefix hso: <http://lipi.go.id/herbarium/> .

hso:2021_03_17_11_52_560001.jpg a hso:Interaction;
hso:hasRegion hso:2021_03_17_11_52_560001.jpg-7739588 .

hso:2021_03_17_11_52_560001.jpg-7739588 a
hso:InteractionMark;
hso:hasRegion _:00106841-cc4a-4674-8851-1a72d4a1a828,
   _:0c688ab1-b924-4301-b0ed-264e73d314a5 .

_:00106841-cc4a-4674-8851-1a72d4a1a828 a hso:Region;
hso:height "362"^^xsd:int;
hso:width "410"^^xsd:int;
hso:x "3606"^^xsd:int;
hso:y "4373"^^xsd:int.
```

```
@prefix hso: <http://lipi.go.id/herbarium/> .

hso:2021_03_17_11_52_560001.jpg a hso:Interaction;
  hso:hasRegion hso:2021_03_17_11_52_560001.jpg-7739588 .

hso:2021_03_17_11_52_560001.jpg-7739588 a hso:InteractionMark;
  hso:hasRegion _:00106841-cc4a-4674-8851-1a72d4a1a828,
                _:0c688ab1-b924-4301-b0ed-264e73d314a5 .

_:00106841-cc4a-4674-8851-1a72d4a1a828 a hso:Region;
  hso:height "362"^^xsd:int;
  hso:width "410"^^xsd:int;
  hso:x "3606"^^xsd:int;
  hso:y "4373"^^xsd:int .

_:0c688ab1-b924-4301-b0ed-264e73d314a5 a hso:Region;
  hso:height "197"^^xsd:int;
  hso:width "330"^^xsd:int;
  hso:x "1151"^^xsd:int;
  hso:y "3260"^^xsd:int .
```

FIGURE 8.6 A snapshot of our annotation.

TABLE 8.3
Produced Annotation

No.	URL	# Triples
1	http://lipi.go.id/herbarium/Interaction	246
2	http://lipi.go.id/herbarium/InteractionMark	246
3	http://lipi.go.id/herbarium/Region	3,396
4	http://lipi.go.id/herbarium/hasRegion	3,642
5	http://lipi.go.id/herbarium/x	3,382
6	http://lipi.go.id/herbarium/y	3,382
7	http://lipi.go.id/herbarium/width	3,382
8	http://lipi.go.id/herbarium/height	3,382

```
_:0c688ab1-b924-4301-b0ed-264e73d314a5 a hso:Region;
hso:height “197”^^xsd:int;
hso:width “330”^^xsd:int;
hso:x “1151”^^xsd:int;
hso:y “3260”^^xsd:int.
```

8.4.4 Discussion

We have introduced a solution to annotate images of herbarium specimens semantically. The annotation can be used as data training for herbivory classification tasks. Unfortunately, most of the existing tools perform the data labeling process individually and have its own format, which is different from the others.

As the consequence, finding a common technique for sharing annotations from one tool to another is difficult. In this paper, we demonstrated how our strategy may solve the annotation discrepancy and become the bridge for multiple tools. Furthermore, the relationship between objects inside annotated specimens is taken into account by our system. As a result, the generated annotation can recognize the objects in the specimens as well as their relationships.

Our schema represented the processes (i.e., the interaction between organisms) as entities. In this case, objects are integrated with the processes, where a process consumes inputs (i.e., parts of a plant) and produced output (i.e., interaction marks such as damages on leaves). This approach is similar to other modeling approaches in multiple domains. For example, the General Formal Ontology [45] in biological and biomedical areas, the OntoDM [46], and the Data Mining Optimization Ontology (DMOP) [47] for data mining processes. A biological interaction was viewed as a process, with actors (such as herbivory) performing actions (such as consuming the part of plant) and cause something (i.e., damages on parts of the plant). It is also important to mention that a set of processes is linked to spatial and temporal data. Each specimen contains the location where the specimen was collected for the spatial information. Objects within specimen images also include the region information where they are found. For the temporal information, each specimen also contains the time when it was collected. Furthermore, performed acts should be described in terms of when they occurred as points in time. Interactions should be distinguishable based on their places and time references. Moreover, as our annotation focused on multiple objects of interest on images, most of them are presented as nested objects. Therefore, we believe that preserving a unique identity for each item is essential for mapping definition. Instead of identifying objects by their position (such as index of an array), it will be better to have an attached identification scheme for consistency throughout the mapping process. This object of interest identification approach would make it easier to keep track of relationships between herbarium specimens, plant parts, and objects of interest contained inside the parts of plant.

8.5 CONCLUSION

Herbarium specimens have become the primary data source for biodiversity research. Multiple organizations collected specimens from various locations and kept them in herbaria all around the globe. The attempt to digitize specimens and share them publicly has piqued the interest of scientific communities, allowing scientists from all around the world to analyze them. As a result, millions of digitized herbarium specimens are available online. From this digital data collection, images are the primary data, accompanied by labels on the specimens (such as taxonomy, spatial information about where the specimens were collected, temporal information about when the specimens were collected, the person who has collected the specimen).

A herbarium specimen holds great valuable information, including spatial and temporal information about the specimen, and other additional information that

can be found from it. For example, plant structures (such as the shape of leaves, and stems) are characteristics that can be extracted from images of herbarium specimens. This type of information can be used to develop intelligent applications, such as a computer vision-based application for automatic species identification. More than that, images of the specimen could also hold the interaction between plant and animal as indicated by the mark of damages or defense mechanisms found on the specimen. The latter type of information can be used further for advanced analytics, such as analyzing invasive species, and global warning indicators.

When the number of digitized herbarium specimens grows exponentially, scientists optimize the data analysis process by automating most of the steps. Artificial intelligence techniques such as machine learning are one option to make it happen. Machine learning techniques, especially supervised ones, require data training to discover the patterns from the data and use them to perform data classification tasks. In this case, machine learning algorithms would use the pattern to classify unknown data. It is widely known that a machine learning algorithm needs a sufficient amount of data training with high quality to produce the best model with highly accurate results. Unfortunately, this kind of data is not always publicly available. Multiple labeling technologies were used to create the majority of the shared data. Therefore, the challenge has shifted from data acquisition to labeling data in cases when there is a label discrepancy.

This work proposes a method to produce high-quality digitized herbarium specimens using semantic annotations. Annotations will be used to identify objects of interest in images and how they are related to one another. The annotation was achieved by employing an ontology to uniformly represent labels of images in a consistent way that is aligned with the goal of any classification tasks at hand. We started by identifying entities found in herbarium specimens before defining relations among them. As a result, the constructed ontology can uniformly represent objects of interest in digitized herbarium specimens. After that, we aligned the ontology with labels generated by multiple image labeling tools through declarative mapping rules. As a result, annotations from digitized herbarium specimens were obtained.

We evaluated our proposed method for an herbivory classification task, where images were labeled with three pre-defined classes. During the mapping process, we discovered that annotations were successfully created with only minimum pre-processing. The main goal of the evaluation was to investigate if we could extract data for machine learning tasks while maintaining the links between objects in the annotation that needed to be semantically represented. Furthermore, by using a shared ontology and declarative mapping rules, we can accommodate a variety of categorization tasks. This work is our first attempt to encode knowledge into machine learning workflows, which remains under-investigated to the best of our knowledge. Moreover, this work is another endeavor to contribute to big biodiversity data management and foster research in this area to move forward faster. In the future, we would like to extend our work by including numerous types of annotation in a variety of categorization tasks across diverse domains.

ACKNOWLEDGMENT

This research was supported by the Research Organization for Life Sciences, National Research and Innovation Agency, Indonesia under the national research priority program "Exploration and Utilization of National Biodiversity", the fiscal year 2021. We would like to thank all members of the Knowledge Engineering Research Group, Research Center for Informatics and the Plant Systematic Research Group, Research Center for Biology for their valuable suggestions and feedback.

REFERENCES

1. S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: A new paradigm to machine learning," *Arch. Comput. Methods Eng.*, vol. 27, no. 4, pp. 1071–1092, Sep. 2020. doi:10.1007/s11831-019-09344-w.
2. J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 67, May 2016. doi:10.1186/s13634-016-0355-x.
3. P. S. Soltis, G. Nelson, A. Zare, and E. K. Meineke, "Plants meet machines: Prospects in machine learning for plant biology," *Appl. Plant Sci.*, vol. 8, no. 6, Jun. 2020. doi:10.1002/aps3.11371.
4. E. K. Meineke, C. C. Davis, and T. J. Davies, "The unrealized potential of herbaria for global change biology," *Ecol. Monogr.*, vol. 88, no. 4, pp. 505–525, Nov. 2018. doi:10.1002/ecm.1307.
5. J. Wäldchen and P. Mäder, "Machine learning for image based species identification," *Methods Ecol. Evol.*, vol. 9, no. 11, pp. 2216–2225, Nov. 2018. doi:10.1111/2041-210X.13075.
6. K. D. Pearson et al., "Machine learning using digitized herbarium specimens to advance phenological research," *Bioscience*, vol. 70, no. 7, pp. 610–620, Jul. 2020. doi:10.1093/biosci/biaa044.
7. W. N. Weaver, J. Ng, and R. G. Laport, "LeafMachine: Using machine learning to automate leaf trait extraction from digitized herbarium specimens," *Appl. Plant Sci.*, vol. 8, no. 6, Jun. 2020. doi:10.1002/aps3.11367.
8. E. K. Meineke, A. T. Classen, N. J. Sanders, and T. Jonathan Davies, "Herbarium specimens reveal increasing herbivory over the past century," *J. Ecol.*, vol. 107, no. 1, pp. 105–117, Jan. 2019. doi:10.1111/1365-2745.13057.
9. C. Beaulieu, C. Lavoie, and R. Proulx, "Bookkeeping of insect herbivory trends in herbarium specimens of purple loosestrife (Lythrum salicaria)," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 374, no. 1763, p. 20170398, Jan. 2019. doi:10.1098/rstb.2017.0398.
10. E. K. Meineke and T. J. Davies, "Museum specimens provide novel insights into changing plant–herbivore interactions," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 374, no. 1763, p. 20170393, Jan. 2019. doi:10.1098/rstb.2017.0393.
11. E. K. Meineke, C. Tomasi, S. Yuan, and K. M. Pryer, "Applying machine learning to investigate long-term insect–plant interactions preserved on digitized herbarium specimens," *Appl. Plant Sci.*, vol. 8, no. 6, Jun. 2020. doi:10.1002/aps3.11369.
12. A. E. White, R. B. Dikow, M. Baugh, A. Jenkins, and P. B. Frandsen, "Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning," *Appl. Plant Sci.*, vol. 8, no. 6, Jun. 2020. doi:10.1002/aps3.11352.

13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. doi:10.1145/3065386.
14. N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?," *PLOS Comput. Biol.*, vol. 4, no. 1, pp. 1–6, 2008. doi:10.1371/journal.pcbi.0040027.
15. N. Zhou et al., "Crowdsourcing image analysis for plant phenomics to generate ground truth data for machine learning," *PLOS Comput. Biol.*, vol. 14, no. 7, pp. 1–16, 2018. doi:10.1371/journal.pcbi.1006337.
16. Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data – AI integration perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1328–1347, Apr. 2021. doi:10.1109/TKDE.2019.2946162.
17. C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, Dec. 2019. doi:10.1186/s40537-019-0197-0.
18. T. Ott, C. Palm, R. Vogt, and C. Oberprieler, "GinJinn: An object-detection pipeline for automated feature extraction from herbarium specimens," *Appl. Plant Sci.*, vol. 8, no. 6, Jun. 2020. doi:10.1002/aps3.11351.
19. B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 157–173, May 2008. doi:10.1007/s11263-007-0090-8.
20. A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2276–2279. doi:10.1145/3343031.3350535.
21. D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, p. 101759, 2020. doi:10.1016/j.media.2020.101759.
22. R. Kays, W. J. McShea, and M. Wikelski, "Born-digital biodiversity data: Millions and billions," *Divers. Distrib.*, vol. 26, no. 5, pp. 644–648, May 2020. doi:10.1111/ddi.12993.
23. S. S. Farley, A. Dawson, S. J. Goring, and J. W. Williams, "Situating ecology as a big-data science: Current advances, challenges, and solutions," *Bioscience*, vol. 68, no. 8, pp. 563–576, Aug. 2018. doi:10.1093/biosci/biy068.
24. R. L. Walls et al., "Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies," *PLoS One*, vol. 9, no. 3, p. e89606, Mar. 2014. doi:10.1371/journal.pone.0089606.
25. R. L. Walls et al., "Ontologies as integrative tools for plant science," *Am. J. Bot.*, vol. 99, no. 8, pp. 1263–1275, Aug. 2012. doi:10.3732/ajb.1200222.
26. L. Cooper et al., "The plant ontology as a tool for comparative plant anatomy and genomic analyses," *Plant Cell Physiol.*, vol. 54, no. 2, p. e1–e1, Feb. 2013. doi:10.1093/pcp/pcs163.
27. Z. Akbar et al., "An ontology-driven personalized faceted search for exploring knowledge bases of capsicum," *Futur. Internet*, vol. 13, no. 7, pp. 1–17, 2021. doi:10.3390/fi13070172.
28. J. Wieczorek et al., "Darwin Core: An evolving community-developed biodiversity data standard," *PLoS One*, vol. 7, no. 1, p. e29715, Jan. 2012. doi:10.1371/journal.pone.0029715.
29. Q. Groom et al., "Improving Darwin Core for research and management of alien species," *Biodivers. Inf. Sci. Stand.*, vol. 3, p. e38084, Oct. 2019. doi:10.3897/biss.3.38084.

30. Z. Akbar, Y. A. Kartika, D. Ridwan Saleh, H. F. Mustika, and L. Parningotan Manik, "On using declarative generation rules to deliver linked biodiversity data," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, 2020, pp. 267–272. doi:10.1109/ICRAMET51080.2020.9298573.
31. R. Hyam, "Semantically linking specimens and images," *Biodivers. Inf. Sci. Stand.*, vol. 3, p. e35343, Jun. 2019. doi:10.3897/biss.3.35343.
32. D. Röpert, F. Reimeier, J. Holetschek, and A. Güntsch, "Semantic annotation of botanical collection data," *Biodivers. Inf. Sci. Stand.*, vol. 3, p. e36187, Jun. 2019. doi:10.3897/biss.3.36187.
33. S. M. Rashid et al., "The semantic data dictionary – An approach for describing and annotating data," *Data Intell.*, vol. 2, no. 4, pp. 443–486, Oct. 2020. doi:10.1162/dint_a_00058.
34. A. Lücking, C. Driller, M. Stoeckel, G. Abrami, A. Pachzelt, and A. Mehler, "Multiple annotation for biodiversity: Developing an annotation framework among biology, linguistics and text technology," *Lang. Resour. Eval.*, Aug. 2021. doi:10.1007/s10579-021-09553-5.
35. A. Kirchhoff et al., "Toward a service-based workflow for automated information extraction from herbarium specimens," *Database*, vol. 2018, 2018. doi:10.1093/database/bay103.
36. L. Stork et al., "Semantic annotation of natural history collections," *J. Web Semant.*, vol. 59, p. 100462, 2019. doi:10.1016/j.websem.2018.06.002.
37. D. Langenkämper, M. Zurowietz, T. Schoening, and T. W. Nattkemper, "BIIGLE 2.0: Browsing and annotating large marine image collections," *Front. Mar. Sci.*, vol. 4, p. 83, 2017. doi:10.3389/fmars.2017.00083.
38. G. S. Mai, F. C. Yang, and M.-N. Tuanmu, "Annotating out the way to the linked biodiversity data web," *Biodivers. Inf. Sci. Stand.*, vol. 1, p. e20270, 2017. doi:10.3897/tdwgproceedings.1.20270.
39. E. W. Schupp, P. Jordano, and J. M. Gómez, "A general framework for effectiveness concepts in mutualisms," *Ecol. Lett.*, vol. 20, no. 5, pp. 577–590, May 2017. doi:10.1111/ele.12764.
40. P. Jordano, "The biodiversity of ecological interactions: Challenges for recording and documenting the Web of Life," *Biodivers. Inf. Sci. Stand.*, vol. 5, p. e75564, Sep. 2021. doi:10.3897/biss.5.75564.
41. A. Iglesias-Molina and D. Chaves-Fraga, "Towards the definition of a language-independent mapping template for knowledge graph creation," Nov. 2019. doi:10.5281/ZENODO.3526141.
42. A. Dimou, M. Vander Sande, and P. Colpaert, "RML: A generic language for integrated RDF mappings of heterogeneous data," in *Workshop on Linked Data on the Web*, Apr. 2014, p. 5.
43. M. A. Musen, "The Protégé project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015. doi:10.1145/2757001.2757003.
44. T. Delva, D. Van Assche, P. Heyvaert, B. De Meester, and A. Dimou, "Integrating nested data into knowledge graphs with RML fields," in *Proceedings of the 2nd International Workshop on Knowledge Graph Construction (KGCW 2021)*, 2021, vol. 2873, p. 16, [Online]. Available: http://ceur-ws.org/Vol-2873/paper9.pdf.
45. H. Herre, "General formal ontology (GFO): A foundational ontology for conceptual modelling," in *Theory and Applications of Ontology: Computer Applications*, R. Poli, M. Healy, and A. Kameas, Eds. Dordrecht: Springer Netherlands, 2010, pp. 297–345.

46. P. Panov, L. Soldatova, and S. Džeroski, "Ontology of core data mining entities," *Data Min. Knowl. Discov.*, vol. 28, no. 5, pp. 1222–1265, Sep. 2014. doi:10.1007/s10618-014-0363-0.
47. C. M. Keet et al., "The data mining optimization ontology," *J. Web Semant.*, vol. 32, pp. 43–53, 2015. doi:10.1016/j.websem.2015.01.001.

NOTES

1 Index Herbariorum, http://sweetgum.nybg.org/science/ih/, Accessed: 05/11/2021.
2 https://www.idigbio.org/portal/, Accessed: 05/11/2021.
3 http://labelme2.csail.mit.edu/, Accessed: 05/11/2021.
4 https://www.robots.ox.ac.uk/~vgg/software/via/, Accessed: 05/11/2021.
5 https://cocodataset.org/\#format-data, Accessed: 05/11/2021.
6 http://host.robots.ox.ac.uk:8080/pascal/VOC/, Accessed: 05/11/2021.
7 https://ricover.hpc.lipi.go.id/ontocapsicum/, Accessed: 08/11/2021.
8 https://www.w3.org/TR/r2rml/, Accessed: 11/11/2021.
9 https://rml.io/, Accessed: 11/11/2021.
10 http://www.biologi.lipi.go.id/botani/index.php/about-bo, Accessed: 06/11/2021.
11 https://ricover.hpc.lipi.go.id/ontocapsicum/, Accessed:14/11/2021.
12 https://fno.io/, Accessed: 20/11/2021.
13 https://github.com/RMLio/rmlmapper-java, Accessed: 20/11/2021.