Accuracy Assessment of Crowdsourced Data in Biological Specimen Transcription

Manash Shah

Master’s Thesis

Faculty of Science and Forestry
Computer Science
June 2014
Abstract

Biodiversity informatics is at a juncture where public engagement in the science is inevitable. Crowdsourcing has emerged as a solution with the possibilities of utilizing intellectual resources of the mass for scientific purposes.

In this study, the possibility of conducting transcription of primary biodiversity data using crowdsourcing has been explored. A crowdsourcing application for transcribing image labels from biological specimens was implemented and the accuracy of transcription from citizen scientists was assessed. The consensus among the users’ inputs was computed using Multiple Sequence Algorithms and then evaluated against the ground truth.

The results demonstrated a mixed level of accuracy for different input fields in transcription. It indicates that the level of accuracy in transcriptions depends on the level of interpretation required from the participants. Further, it indicates that implementation of suggestive features in the application improves accuracy.

Keywords: Biodiversity Informatics, Darwin Core, Crowdsourcing, PyBossa, Multiple Sequence Alignment
Preface

I would like to express my gratitude to the University of Eastern Finland and the School of Computing for giving me this invaluable opportunity. I am extremely thankful to Prof. Markku Hauta-Kasari for his guidance and supervision. I am deeply grateful to my colleague and mentor Tero Mononen without whom the study would not have been possible. I am indebted to PhD, Riitta Tegelberg and PhD, Hannu Saarenmaa, Digitarium, University of Eastern Finland for their help, encouragement and guidance. I would like to convey my heartfelt thanks to Paras Pant and PhD, Ville Hautamäki for their valuable comments, suggestions and opinions. I also like to express my appreciation for Tim Robertson, GBIF Secretariat and PhD, Daniel Lombraña González for their prompt responses and help whenever in need. I owe my sincere thanks to all the participants of the crowdsourcing task.

Finally, I would like to devote my sincerest gratitude to my friends and family for the unconditional love and support they have blessed me with. John Ronald Reuel Tolkien in his poem rightly remarked “Not all those who wander are lost”. My deepest gratitude to all of you who have been a guiding light in my pursuit.

Joensuu, June 2014

Manash Shah
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>UEF</td>
<td>University of Eastern Finland</td>
</tr>
<tr>
<td>TDWG</td>
<td>Taxonomic Databases Working Group</td>
</tr>
<tr>
<td>GBIF</td>
<td>Global Biodiversity Information Facility</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
<tr>
<td>URI</td>
<td>Uniform Resource Identifier</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma-Separated Values</td>
</tr>
<tr>
<td>ABCD</td>
<td>Access to Biological Collections Data</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple Sequence Algorithm</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>REST</td>
<td>Representational State Transfer</td>
</tr>
<tr>
<td>HTML</td>
<td>HyperText Markup Language</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>CSS</td>
<td>Cascading Style Sheets</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>MAffT</td>
<td>Multiple Alignment using Fast Fourier Transform</td>
</tr>
</tbody>
</table>
5.3 Multiple Sequence Alignment ........................................... 37
  5.3.1 FASTA format ......................................................... 38
  5.3.2 Multiple Alignment using Fast Fourier Transform (MAFFT) . 39
5.4 Outliers Removal ......................................................... 41
5.5 Consensus ................................................................. 42

6 Results ........................................................................... 44

7 Conclusion and Recommendations ........................................ 49

Bibliography ........................................................................ 51

Appendix A A generic schema of Darwin Core terms .................. 59

Appendix B Darwin Core classes ............................................ 69
1 Introduction

Biodiversity Informatics is an interdisciplinary field encompassing biological studies with information technology. It is a young and evolving science that implements and utilizes concepts or methods from computer science, mathematics and statistics to biological diversity studies and data [1]. The study fosters us with an improved understanding about the flora and fauna of the planet [2, 3].

Crowdsourcing is an emerging concept of harvesting the collective human intelligence for attainment of set goals through the voluntary participation of users of diverse nature in the endeavor via an online system that surpasses the traditional boundaries which were otherwise impassable[4, 5].

The study is an amalgamation of ideas from different discipline and an exploration of methods and solutions for analyzing accuracy of biodiversity data gathered through crowdsourcing application using mathematical and algorithmic techniques that could contribute to the enhancement of current practices and processes.

1.1 Motivation

The efforts in the study of biodiversity and the primary biodiversity data have witnessed exponential growth both in terms of volume and quality in the recent years. The realization of the importance of biodiversity studies and the emergence of affordable technologies have revolutionized the digitization process in biodiversity studies [1, 6]. Limited by the availability and efficiency of humans, the human intelligence aspect brought in by the experts of the field that is irreplaceable even by the mighty machine, lags behind creating a backlog in the digitization process [7, 8, 9]. Crowdsourcing the parts of the process that require human cognition could be a possible solution as it can reduce the cost incurred in experts’ involvement and increases the efficiency at the same time [5]. However, assessing and maintaining the quality of data generated from crowdsourcing endeavors is a herculean challenge. The exploration of the possibilities and process of addressing the aforementioned challenge is an active field of research and motivation for the undertaking of current study.
1.2 Purpose

The digitization of natural history collections has evolved from a manual process to an industrial automated process. Digitarium, the digitization center at the University of Eastern Finland has implemented an automated conveyor imaging system with a throughput of two thousand images per day [7]. However, transcribing the labels on the image is still manually conducted by the experts at the facility [10]. The number of biological specimens imaged outnumbers the number of transcription that can be performed by the limited number of available experts. This inconsistency has created a backlog and therefore a sustainable solution is of utmost necessity.

The study had a twofold purpose, the first being the design and implementation of a crowdsourcing application for transcription of biological specimens. The objective of the study was to implement a crowdsourcing platform where volunteers could participate in transcription of labels on images of biological specimens. This application further served the requirement of the study acting as a means of collecting data that could be assessed against the ground data.

The second purpose of the study was to implement a workflow or technique of generating consensus among the data collected from the crowdsourcing application that could be evaluated against the ground data. The objective of the study was to implement a workflow for building the consensus outlined by a similar scientific study [11, 12] and assess its validity in the present context.

1.3 Organization

The thesis comprises seven chapters with the first being the current introductory chapter that presents an overview of the subject of the study, the motivation and the purpose of conducting the study.

The second chapter introduces Biodiversity informatics, primary biodiversity data and elaborates its importance. The chapter further delves into digitization of biodiversity data, the challenges and the current state of digitization. The chapter also presents the data standard used in biodiversity data exchange, the Darwin Core standard.

The third chapter presents Crowdsourcing and various aspects pertaining to it. The
chapter presents the definition of the concept and the components encompassing the process.

The fourth chapter describes PyBossa, the open-source crowdsourcing framework used for creating the application used in the current study. The chapter briefly presents the architecture of the framework. The chapter further describes the crowdsourcing application implemented for transcription of biological specimens in the study.

The fifth chapter details the process involved in data analysis and building consensus among data from crowdsourcing. It presents the workflow, algorithms and the methods in pre and post processing of data along with the data format used in the study.

The results of the study along with the probable explanations are presented in the penultimate chapter, the sixth chapter.

The final chapter presents findings of the study and suggests probable recommendations for further improvements and study.
2 Biodiversity Informatics

Biodiversity, a modern contraction of the term Biological diversity, is an umbrella term that refers to the variety of life on earth in all its forms. The term was apparently coined by Walter G. Rosen in 1986 and most likely first appeared in print literature in 1988 [13]. The Convention on Biological Diversity [14] defines it as “the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems”. Our understanding of biodiversity is of utmost value for formulation of policies and decisions on issues like conservation, land-use, environmental change, sustainability, food security and ecosystem [3].

Biodiversity informatics is defined as application of informatics technique for management, algorithmic exploration, analysis and interpretation of primary data regarding biodiversity [1]. It is an emerging realm that broadly covers the creation, refinement, analysis, interpretation of raw data from observation of biodiversity over space and time along with ecological data [15]. It fosters us with the much-needed information to address the pressing issues concerning environment and ecosystem.

The study of biological diversity has been in practice for decades and with the Convention on Biological Diversity that entered into force in December 1993, the importance of study of biodiversity was realized and established. The convention laid down its three main objectives as the conservation, sustainable use and sharing of the benefits from the utilization of resources [14]. All the three objectives require our enhanced understanding of biological diversity and hence the study of biodiversity was envisioned and accepted to be of utmost importance both by scientific community and policymakers alike. With the advent of Information technology in the recent years, the process and practices in the study of biological diversity has undergone transformation embracing the ubiquitous digital revolution to evolve into a new discipline of Biodiversity Informatics [16]. The application of informatics tools and techniques on primary data to create digital objects and information has enabled exchange and integration of biodiversity data from collections and observations around the world. This has further taken the exploration of world biodiversity to an unprecedented level where researchers are now able to model and study biodiversity patterns, processes and impact at both the macro and micro-scale [15].
2.0.1 Biological Specimen

Biological Specimen refers to the biological material preserved as sample representing a particular species for study and research. Natural history collections are the repositories of primary data on biodiversity and are home to millions of biological specimens that serve as the window to past, present and the future of flora and fauna of the planet. The collections provide primary observation data of species across the planet in course of history along with information on the habitat, geography and other features related to their occurrence [1].

2.0.2 Importance

Natural history collections have a broad range of use. Accurate identifications of species require voucher specimens for corroboration. In some cases, the specimens stored in the natural history museums are our only links to the extinct species. Data from historical collections can be used as baseline data against which the present observations can be vetted to assess the changes in biodiversity pattern [2, 17].

The data from collections are vital for education, both formal and public. Studies on taxonomy, biogeography, species diversity, population rely heavily on collections data. Collections in museums provide the ground data for determination, description and classification of new and existing taxa alike [18]. The collections are the basic source for understanding the changes in landscape and patterns of biodiversity distribution [2, 14, 17]. They are indispensable in the formulation of species distribution atlases, for species distribution modeling and for prediction of new species distributions. Primary collection data are necessary for understanding of species declines over space and time and for monitoring current species numbers for prevention of future declines [19]. The study of species richness, density and identification of centers of endemism are immensely aided by species occurrence data. Studying individual species richness or its evolutionary patterns cannot be completed without the primary species data [2, 14]. Primary occurrence data is needed for modeling of species population and for understanding the dynamics of the population.

Further, the data fuel the information necessary for understanding of life history and phenology [18]. The various stages of lifecycle and its relationship with geographic locations or time can be studied using primary data [2]. Phenology is the study of
the timing of naturally occurring events and their relationship of biotic and abiotic variables [2]. Primary occurrence data are essential to phenological studies.

Primary occurrence data are important resources for studies on endangered, migratory and invasive species. Identification of endangered species, understanding of reasons behind their endangerment and development of strategies for protection cannot be complete without primary data [2, 17]. In cases of invasive species, the identification of its native location, the determination of its niche characteristics and prediction of its likely spread can be conducted using species occurrence data. The availability and exchange of species-occurrence data between museums and researchers from different countries are essential to tracking and monitoring of migratory species possible across political boundaries [2, 16].

Species occurrence data are being used in studies on impact of climate change on native species in a region, on primary production in agriculture or identifying the indicators of desertification [2]. The primary occurrence data are the resource for identifying the pattern, understanding the processes and causes of evolution, ecological phenomena and genetic relationships [2, 16].

Further, these data are required in formulation of environmental regionalization [2]. They provide the relevant information to determine conservation priorities and building of strategies and policies on biodiversity conservation. Information on biodiversity from primary occurrence data aids in the efforts made in management of natural resources; both land and water and further helps in environment protection and monitoring.

Moreover, primary occurrence data provide vital information for sustainable use of resources in agriculture, forestry, fisheries and mining. Primary occurrence data have a huge importance in studies of new species or wild relatives of cultivated crops, their provenances, in food processing and many other agroindustry related studies [2]. Forestry industry requires and utilizes the primary occurrence data for balancing forestry and conservation, in plantation forestry endeavors and in identification of appropriate provenances. Fishing and fisheries use occurrence data for studies on sustainable management of both marine and fresh-water fisheries, controlling overfishing, identification and reduction of bycatch in commercial fishing and monitoring of contamination [2]. Species occurrence data are used in mining industry in identification of high mineral concentrations, in mine site regeneration and in pollution monitoring and
control.

Natural history collections are not just a treasure trove for scientist but have significant utility in art, history and recreation, and for social and political use. They have vital importance in ecotourism and bioprospecting and are the source of attraction to public in awareness towards sustainable development, ecological conservation and endeavors alike [2].

2.1 Biological Specimen Digitization

Biological Specimen Digitization refers to the process of creating digital object and documenting biological specimen along with its related observational records. The process involves digital capture of the physical specimen in appropriate media, viz. visual, audio, including their labels and associated materials along with observational data, georeferencing the associated locality information [16, 10]. The process further extends to appropriate digital archiving and curatorial activities adhering to community-accepted standards for creation of shared and distributed biodiversity data resources that are universally available and accessible.

The purpose of creating information in digital format is to provide users with the ability to query and analyze data in a manner which is both cost-effective and unhindered by geographical and political barriers. Further, this assists in the effort to preserve the rich natural history and diversity for future generations despite the fact that some of it might have been or is in the process of being lost in due course of time.

2.1.1 Challenges

Preliminary estimates indicate 1.2 to 2.1 billion units (specimens, lots and collections) of natural history collection data worldwide [9]. The information that lay scattered in the collections if utilized in a systematic manner can provide us with better insights to cope with the challenges that lay ahead. However, the ground reality is that only a nominal subset is currently accessible to researchers and the general community. The majority of the sample specimens still remain locked in museums quite inaccessible to the scientific communities around the globe due to political and geographical distances. The collaborative task of making such a vast treasure of knowledge accessible to all
the stakeholders in itself is a challenge [16].

The recent leaps and bounds in information and communication technologies have made dissemination of biodiversity data in digital format possible [6, 20]. The creation of digital object from physical specimen has overcome the geographical barriers in exchange and exploration of biodiversity information. The importance of digitization of natural history collections has been realized and appreciable efforts for the same have been initiated.

The process of digitization of natural history has its own challenges and issues that need to be addressed before the knowledge resource from it can be utilized to its potential [8, 21]. The biggest challenge and the most prevalent one is the lack of financial resources or issues related to it. Digitization of natural history collections has not received sufficient priority at national or international arena and hence the prevalent liquidity crunch across almost all the digitization efforts. Quality of data produced from digitization projects of diverse origin, priorities and practices has been identified as another challenge [8].

### 2.1.2 Mass Digitization

The number of biological specimens that remain locked up inside natural history museums around the world are staggering [9, 21, 10, 7]. The invaluable primary biodiversity data the repositories hold have remained inaccessible to most of the scientific community, and the loss is immeasurable [22]. The current pace of digitization would take hundreds of man-years before we can reach anywhere near utilizing the resource to its full potential [7]. The only viable solution to mitigate this loss is to take the digitization process at an industrial scale. The necessity has been realized and investment in research and development of infrastructure to meet this challenge has certainly increased.

A number of appreciable efforts have been made and stories of success have been reported [10, 8, 20]. Dedicated digitization centers with automated imaging system capable of production at industrial efficiencies have been established. Digitarium, the digitization center at the University of Eastern Finland is one such success story with state-of-art infrastructure for mass digitization of biological specimens [7]. The infrastructure and the knowledge base has contributed to the digitization initiatives across
Digitization involves a variety of tasks like identification, annotation, transcription and georeferencing of localities. Novel methods and approaches need to be explored and innovated for these tasks to be conducted in tandem with the imaging or capturing of primary data before they can be disseminated into the digital ecosystem [8, 7]. The tasks involved in the digitization process like identifying, transcribing or georeferencing specimen data from images are being transferred from laboratories to the mass through online systems and crowdsourcing platforms.

Metadata standards for creation and sharing of biodiversity data have been defined and brought into practice [15]. The standards and protocols have laid down the norms and common practice for a collaborative environment. This increased interoperability inducted by such initiative has had positive impact on the level of sharing which eventually has benefited the scientific communities and other stakeholders alike.

### 2.2 Darwin Core Standard

Darwin Core Standard is a data standard for exchange of biodiversity data incorporating a set of vocabulary for exchange of biodiversity information. The standard was drafted and is maintained by Biodiversity Information Standards (TDWG), also known as the Taxonomic Databases Working Group [23, 24]. The current version of Darwin Core standard is based on the standards developed by the Dublin Core Metadata Initiative (DCMI) and has been designed to facilitate systematic discovery, extraction, integration and exchange of primary biodiversity information documented in observations, physical specimens and samples using well-defined standard core vocabulary in a flexible framework [23]. The standard aggregates terms that cover metadata about the specimen, its identification, its taxonomic classification, details of its locality and habitat, information on the collecting event, biological data related to the specimen and other references like digital object identifier or images of the specimen.

#### 2.2.1 Importance

Biodiversity informatics, like any other information technology discipline managing huge amount of data, is plagued by the issue of data quality. The volume of data, diver-
sity of source and the timeline over which the primary data have been collected add to the complexity of streamlining and maintaining the quality. Darwin Core standard is the outcome of an enormous amount of time and effort from the natural history collection community in addressing the issue. Further, it promotes interoperability and facilitates exchange of information over geographical or technological barriers [24, 25].

2.2.2 Description

Darwin Core Standard is maintained in Resource Description Framework (RDF), a standard model for data interchange on the web. The purpose of the standard is to define and maintain a shared but uniform lexicon for use in biodiversity informatics and related domains alike. The standard primarily comprises a vocabulary of terms with clear and concise semantics that is both human and machine readable along with reference definitions, examples and commentaries for consistent use of the terms. A generic schema of Darwin Core terms in eXtensible Markup Language (XML) is detailed in Appendix A.

Further, it includes the namespace policy of the terms which governs the management of the terms for increment in the capabilities of the standard through refinement or addition of new terms. Unique Uniform Resource Identifier (URI) must be assigned to all the terms in Darwin Core for identification and are grouped into collections known as Darwin Core namespaces [23, 24].

The standard further includes the document describing decisions regarding amendment of terms over time. This gives an overview of reasons and results of proposed changes to the terms and standards [23].

A complete specification of the terms and their histories form an integral part of the standard which define and describe the usage of the terms and serve as a guideline for extension of the terms and vocabulary for adaptation to new necessities.

The standard does not make enforcements related to data content but it provide recommendations for best-practices and quality assurance [23, 24]. Recommendations on types, valid values or controlled vocabularies are made in the standard leaving the implementation of such regulations to applications.

The terms are classified into classes viz. Occurrence, MaterialSample, Event, Location,
Identification, Taxon, GeologicalContext, Record-level Terms, ResourceRelationship, MeasurementOrFact. The categories Occurrence, MaterialSample, Event, Location, Identification, Taxon and GeologicalContext consist of terms that relate to the realm of biodiversity and the remaining classes ResourceRelationship, MeasurementOrFact and Record-level Terms cover the relationships to other resources, measurements and generic information about the record [23]. Table 1 tabulates the Darwin Core classes and their respective terms.

Table 1: List of Darwin Core Classes and Terms.

<table>
<thead>
<tr>
<th>Darwin Core Class</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence</td>
<td>occurrenceID, catalogNumber, occurrenceRemarks, recordNumber, recordedBy, individualID, individualCount, sex, lifeStage, reproductiveCondition, behavior, establishmentMeans, occurrenceStatus, preparations, disposition, otherCatalogNumbers, previousIdentifications, associatedMedia, associatedReferences, associatedOccurrences, associatedSequences, associatedTaxa</td>
</tr>
<tr>
<td>MaterialSample</td>
<td>materialSampleID</td>
</tr>
<tr>
<td>Event</td>
<td>eventID, samplingProtocol, samplingEffort, eventDate, eventTime, startDayOfYear, endDayOfYear, year, month, day, verbatimEventDate, habitat, fieldNumber, fieldNotes, eventRemarks</td>
</tr>
<tr>
<td>Location</td>
<td>locationID, higherGeographyID, higherGeography, continent, waterBody, islandGroup, island, country, countryCode, stateProvince, county, municipality, locality, verbatimLocality, verbatimElevation, minimumElevationInMeters, maximumElevationInMeters, verbatimDepth, minimumDepthInMeters, maximumDepthInMeters, minimumDistanceAboveSurfaceInMeters, maximumDistanceAboveSurfaceInMeters, locationAccordingTo, locationRemarks, verbatimCoordinates, verbatimLatitude, verbatimLongitude, verbatimCoordinateSystem, verbatimSRS, decimalLatitude, decimalLongitude, geodeticDatum, coordinateUncertaintyInMeters, coordinatePrecision, pointRadiusSpatialFit, footprintWKT, footprintSRS, footprintSpatialFit, georeferencedBy, georeferencedDate, georeferenceProtocol, georeferenceSources, georeferenceVerificationStatus, georeferenceRemarks</td>
</tr>
<tr>
<td>Identification</td>
<td>identificationID, identifiedBy, dateIdentified, identificationReferences, identificationVerificationStatus, identificationRemarks, identificationQualifier, typeStatus</td>
</tr>
<tr>
<td>Taxon</td>
<td>taxonID, scientificNameID, acceptedNameUsageID, parentNameUsageID, originalNameUsageID, nameAccordingToID, namePublishedInID, taxonConceptID, scientificName, acceptedNameUsage, parentNameUsage, originalNameUsage, nameAccordingTo, namePublishedIn, namePublishedInYear, higherClassification, kingdom, phylum, class, order, family, genus, subgenus, specificEpithet, infraspecificEpithet, taxonRank, verbatimTaxonRank, scientificNameAuthorship, vernacularName, nomenclaturalCode, taxonomicStatus, nomenclaturalStatus, taxonRemarks</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>GeologicalContext</td>
<td>geologicalContextID, earliestEonOrLowestEonothem, latestEonOrHighestEonothem, earliestEraOrLowestErathem, latestEraOrHighestErathem, earliestPeriodOrLowestSystem, latestPeriodOrHighestSystem, earliestEpochOrLowestSeries, latestEpochOrHighestSeries, earliestAgeOrLowestStage, latestAgeOrHighestStage, lowestBiostratigraphicZone, highestBiostratigraphicZone, lithostratigraphicTerms, group, formation, member, bed</td>
</tr>
<tr>
<td>Record-level Terms</td>
<td>dcterms:type, dcterms:modified, dcterms:language, dcterms:rights, dcterms:rightsHolder, dcterms:accessRights, dcterms:bibliographicCitation, dcterms:references, institutionID, collectionID, datasetID, institutionCode, collectionCode, datasetName, ownerInstitutionCode, basisOfRecord, informationWithheld, dataGeneralizations, dynamicProperties</td>
</tr>
<tr>
<td>ResourceRelationship</td>
<td>resourceRelationshipID, resourceId, relatedResourceId, relationshipOfResource, relationshipAccordingTo, relationshipEstablishedDate, relationshipRemarks</td>
</tr>
<tr>
<td>MeasurementOrFact</td>
<td>measurementID, measurementType, measurementValue, measurementAccuracy, measurementUnit, measurementDeterminedDate, measurementDeterminedBy, measurementMethod, measurementRemarks</td>
</tr>
</tbody>
</table>

Darwin Core can be implemented for sharing data in various encoding schemes viz. Comma Separated Values (CSV), eXtensible Markup Language (XML), JavaScript Object Notation (JSON), RDF etc. The standard also includes mapping mechanism to concepts in the Access to Biological Collections Data schema (ABCD). A schema of Darwin Core classes and their respective terms in eXtensible Markup Language (XML) is detailed in Appendix B.
3 Crowdsourcing

Crowdsourcing, a relatively new concept, was coined by Jeff Howe [4] in 2006 to describe the emerging web-based business model of engaging the vast and diverse competence and expertise of a mass of people via an open call. Howe proposes the following two definitions:

The White Paper Version: “Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call” [4].

The Soundbyte Version: “The application of Open Source principles to fields outside of software” [4].

An effort to form an exhaustive and global definition of Crowdsourcing from review, analysis and integration of definitions appearing in scientific literature and digital media has presented an integrated definition of crowdsourcing as follows:

“Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken” [26].

The process of crowdsourcing encompasses three key elements namely crowd, the initiator and the process [27]. The inception of the process is marked with the initiator presenting a task or a set of tasks to the crowd or digital citizen for their voluntary engagement in resolution of the stated problem via an open call on the internet.

The crowd can be described as a group of individuals responding to the initiators call for voluntary participation. The size of the crowd cannot be predicted and is normally heterogeneous in nature from the knowledge, expertise or motivational point of view. The crowd is expected to bring in its time, effort, knowledge and/or creativity for com-
pletion of the presented tasks. The presented tasks in turn might be simple, complex or creative in nature [5] and thus the initiative itself defining the nature of the targeted participants. The motivation of the individuals in the crowd for participation ranges from obtaining compensation of financial nature to personal development or satisfaction from social contribution.

The initiator refers to any entity that initiates the crowdsourcing process seeking solution to their problem which demand huge workforce or situation that necessitate seeking of external expertise or accessing the creativity of individuals [28]. The initiator can be non-profit or for-profit body and hence the reward or remuneration for the contribution can be of intangible nature or tangible recompense ranging from micro payments to huge financial rewards.

The process of crowdsourcing is an online process and inherits the participatory and distributive nature of the internet. It is participatory as the call made is more or less open to general public with participation from the crowd being dictated more by their own capabilities or motivations. The process is primarily conducted on the internet, though there might be the possibility of performing tasks offline, thereby making the process distributive and overcoming political and geographical boundaries in seeking solutions.

### 3.1 Tasks

The tasks presented through a crowdsourcing platform can be of varied nature [29, 30]. It can be a large volume of simple mundane tasks which requires large volume of workforce not available within the premises of the crowdsourcers’ organizational boundaries. In contrast, a unique problem which expects solution with new ideas, innovation or novelty from a heterogeneous group of individuals for challenging complex tasks can also be presented for resolution through crowdsourcing [31]. Further, tasks that requires public involvement for creation or enhancement, artistic opinion, reviews or ratings in fulfilling requirements of creative nature can also be presented via crowdsourcing platform [5].
3.2 Process

The process or paradigm in a Crowdsourcing initiative relies on the nature of the problem. The process can be integrative, selective or a hybrid of the two [29]. Simple tasks of repetitive nature which demand lower cognitive process but require assimilation and analysis of variety of user input to find an aggregate solution might utilize the integrative approach. Complex tasks which are domain specific without known solutions requiring higher expertise and innovation adhere to the selective approach as multiple solution to the problem might not be expected rather choosing optimal solution from a set of candidate solutions might be best suited to the situation. The process in crowdsourcing of creative tasks have both the integrative and selective nature. The selective process brings out the creativity and uniqueness in the solution and the integrative process derives the appeal to the mass [29].

3.3 User Motivation

User participation is a must for any crowdsourcing initiatives although the size of the crowd expected by the initiator depends on the endeavor undertaken [32]. It is one of the biggest challenge in crowdsourcing to intrigue the users to participate and keep them motivated until completion of the tasks. A crowdsourcing setting comes with the drawback of initiator or employers’ lack of control or overview in selection of crowd workers in regard to their capabilities and motivation [29]. This in itself is a bigger threat to achievement of set goals, be it in terms of time or quality of the output. The initiator can in some cases implement screening mechanism for vetting crowd workers’ capabilities, but does not have much control over the motivational factors. Motivation of the masses to voluntarily participate in such activities can be aroused by different stimulus [33].

Prior studies have identified a number of incentives or motivation for the crowd to participate in crowdsourcing work and broadly classified them into two categories intrinsic motivation and extrinsic motivation [34, 35, 33].

A number of aspects of crowdsourcing initiatives and tasks have been identified to act as intrinsic motivator. The diversity of expertise required for participation, the perception of result or effect of the task performed, the degree of freedom and creativity, a
sense of accomplishment through immediate feedback and a productive way of killing boredom are some of the identified personal incentives that appeal the masses toward crowdsourcing [34, 35]. Reputation in the community, peer pressure and networking are some social motives of intrinsic nature that woo the masses. These incentives are intrinsic in nature and are identified to by the crowd at an individual level.

Incentives with immediate payoffs or those with futuristic payoffs or social motivations concerning values, norms and obligations considered significant by external factors are extrinsic motivations for the crowd to participate [28, 33]. Monetary remuneration for the tasks performed is a major incentive for participation [34, 28, 36]. Similarly, participation for displaying presence in the community or for personal advancement and grooming are motivating factors with futuristic values to the crowdworkers [33]. Further, exhibition of solidarity to values deemed important by the initiator, external stakeholders and by the participant act as a strong motive. Likewise, participation to meet ones obligation or participation with the expectation of feedback or commendation have been identified to be strong social incentives of extrinsic nature in crowdsourcing endeavors [28].

One of the most accepted model proposed for user’s motivation in crowdsourcing is illustrated in Figure 1.

![Worker's Motivation in Crowdsourcing model](image)

Figure 1: A Model for Worker’s Motivation in Crowdsourcing proposed in [37].

The relation and effect of different intrinsic and extrinsic motivation factors on accuracy, throughput, quality and similar other success measures of a crowdsourcing project has been a subject of study [34, 38]. There seem to be a unanimity on the notion that improvement of monetary remuneration increases the throughput of the project. Likewise, studies indicate that intrinsic motivator have a direct relationship to accuracy
of the output especially in situations where extrinsic motivational factors have lower impact [34, 39].

### 3.4 Crowdsourcing and Biodiversity Informatics

Crowdsourcing can be employed in situations where a bigger task can be divided into smaller tasks and disseminated into the mass for completion. It is equally useful for tasks of lower complexity but requiring human cognition. The transcription of data from labels on biological specimens is a suitable task for crowdsourcing [40]. The volume of work that need to be done in a mass digitization project also necessitates the task-force that can be obtained through crowdsourcing [7, 28]. The nature of sub-tasks like transcribing image labels require human intelligence at a larger number, rather than a higher level. This makes reaching out to an unknown group of people of diverse background to participate in the process possible.

A number of crowdsourcing application have been implemented for transcription of image labels on biological specimens [41, 42, 43]. However, the quality of the data collected is always a high priority in crowdsourcing. During the current study, a crowdsourcing platform for voluntary participation for transcription of image labels is implemented, alongside an exploratory study to evaluate the quality of the crowdsourced data.
4 Implementation

Data Collection for the study was conducted through a crowdsourcing application implemented on PyBossa framework [44]. The application for data collection was implemented and hosted on the server at Digitarium, University of Eastern Finland and was accessible via “http://manash.eu/”. Citizen scientists were requested for voluntary participation via mass email and personal messages. The application was hosted for a period of three months during which more than thirty users registered for participation and volunteered for the tasks assigned. The framework employed, the application designed and deployed, the tasks and the relevant processes in data collection shall be detailed in the subsequent chapters.

4.1 PyBossa

PyBossa is an open source framework for crowdsourcing implemented using Python programming language which can be deployed as a platform for creating microtasking web applications for volunteer participation in tasks requiring human cognition, knowledge or intelligence [44]. PyBossa in itself is not a crowdsourcing application but a framework for developing and deploying crowdsourcing applications. The framework has been designed for volunteer-driven projects and does not have the features to manage financial reimbursement.

4.1.1 PyBossa Architecture

PyBossa consist of two main components, the server and the application. The server is a python web application designed to manage tasks for the calling web applications and is inspired and derived from the BOSSA architecture [45]. PyBossa server provides a platform for creation and presentation of applications and their corresponding tasks. Figure 2 illustrates an overview of the functioning of a PyBossa system.
Applications can be created in PyBossa using the web interface or the RESTful API. The applications hosted on a PyBossa server essentially consist of two parts namely the Task Creator and the Task Presentor.

Every application in PyBossa requires a Task Creator for creating Tasks and related entities. The task creator is a python script which basically operates externally and interacts with the PyBossa core via the API. The task creator script uploads the application tasks to the PyBossa server. The tasks can be uploaded into PyBossa in custom defined spreadsheet templates, CSV or online spreadsheet.

The front-end activities or the end user’s interactions are handled by the Task Presenter. It is a dynamic HTML user interface powered by JavaScript functions that load the tasks and accept user responses. The task presenter can be customized to the tasks that the application presents and to suit the format of the response from the user. The task presenter can be customized to load simple forms, images, maps, sound, PDF files, etc. using JavaScript libraries as per the needs of the application implemented. The user response can be formatted and handled to be accommodated in JSON format supported by the PyBossa server.
4.1.2 PyBossa Domain Model

PyBossa consists of five domain objects namely App, Task, TaskRun, User and Category [47]. The domain objects are reflected in the PyBossa database and are accessible via a RESTful API. Table 2 summarizes the PyBossa domain objects.

<table>
<thead>
<tr>
<th>Domain Object</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>App</td>
<td>A PyBossa application which has tasks associated to it. The application is assigned to a category.</td>
</tr>
<tr>
<td>Task</td>
<td>An individual task associated to an App that can be performed by a participant.</td>
</tr>
<tr>
<td>TaskRun</td>
<td>The response of a specific user’s participation in a specific task. Each object is associated to a Task and a User object.</td>
</tr>
<tr>
<td>User</td>
<td>A participant or user account.</td>
</tr>
<tr>
<td>Category</td>
<td>A category to which applications are associated to.</td>
</tr>
</tbody>
</table>

PyBossa implements a simple domain model to represent the crowdsourcing workflow. Figure 3 illustrates the PyBossa domain objects and their interrelationships.

![Figure 3: PyBossa Domain Model [47]](image-url)
4.1.3 PyBossa RESTful API

PyBossa server can host multiple applications and deliver tasks through a RESTful API which expects and returns JSON encoded requests and responses [48]. The API supports the following operations listed in the Table 3.

Table 3: PyBossa RESTful API supported operations [48].

<table>
<thead>
<tr>
<th>Operation</th>
<th>Action/Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>List</td>
<td>List Domain Objects.</td>
</tr>
<tr>
<td>Get</td>
<td>Get a specific domain object by id. Returns domain object.</td>
</tr>
<tr>
<td>Search</td>
<td>Get a list of domain objects by its fields. Returns a list of domain objects matching the query.</td>
</tr>
<tr>
<td>Create</td>
<td>Create a domain object. Returns created domain object.</td>
</tr>
<tr>
<td>Update</td>
<td>Update a domain object.</td>
</tr>
<tr>
<td>Delete</td>
<td>Delete a domain object.</td>
</tr>
<tr>
<td>Requesting a new task</td>
<td>Request a new task for the current user (anonymous or authenticated).</td>
</tr>
<tr>
<td>Requesting user’s oAuth tokens</td>
<td>Request his/her own oAuth tokens by user registered with any of the third parties supported by PyBossa.</td>
</tr>
</tbody>
</table>

4.1.4 Features

PyBossa incorporates various features that make the framework and the deployed applications easy to use, store data, manage users, workflows and processes [49]. Anonymous participation, participation by registration and registration through third party authentication are supported in PyBossa which can be customized by the initiator depending on his/her requirements. PyBossa has functionalities to provide statistical information on applications, tasks and users alike. The framework has the features to upload tasks in different formats like spreadsheet templates, CSV or online spreadsheet. Similarly, user responses can be exported in CSV or JSON format as per the
requirement.

4.2 Biological Specimen Transcription Application

4.2.1 Tasks

A set of 100 herbarium specimen images were chosen randomly for transcription through volunteer participation. The selected herbarium specimen had been transcribed and verified by experts at Digitarium, University of Eastern Finland and hence could be used as ground data for validation. The set consisted of specimens with both hand written labels and printed labels alike. Some of the specimens were labeled in Finnish language while some in English. Hence the chosen specimen presented diversity in terms of readability along with the difficulty of interpretation.

The set of 100 specimen where further divided into 5 subsets of 20 specimen each and created as 5 different applications each containing 20 tasks. The presentation of a set of specimens for transcription in smaller subsets as separate applications was done in such a fashion so that the motivation level of the volunteer participating in the tasks did not decrease from the volume of work to be done. The division into subsets allowed the participant to achieve a sense of accomplishment after completion of a smaller number of tasks and helped improve their participation and higher engagement.

Each transcription task consisted of identifying and transcribing nine different data fields. The nine fields selected for transcription were the fields that appeared most frequently on the specimen labels and required lower levels of interpretation from the transcriber. Such a selection was made based on discussions and suggestions received from the experts who had actively participated in transcription of the chosen set and similar specimens. The data fields presented for transcription and their corresponding Darwin Core terms are tabulated in the Table 4.

<table>
<thead>
<tr>
<th>Data Fields</th>
<th>Darwin Core Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific Name</td>
<td>scientificName</td>
</tr>
<tr>
<td>Scientific Name authorship</td>
<td>scientificNameAuthorship</td>
</tr>
<tr>
<td>Collected by</td>
<td>recordedBy</td>
</tr>
</tbody>
</table>

Table 4: Data Fields for Transcription
Each image or task was presented for transcription to five users. A task was marked completed when it was transcribed by five unique users. This repetition was performed to gather data such that a consensus could be drawn between the responses of the users. The process, model and results of the consensus from users’ input shall be discussed in subsequent chapters.

4.2.2 User Interface

The user interface implemented for participation was done using HTML, external JavaScript and CSS libraries. The interface had a simplistic layout based on responsive web design approach using BootStrap design framework [50]. An external library was employed for viewing and handling of images. The JavaScript library, jQuery [51] was used for the necessary document traversal, manipulation, event handling and data related processes. Figure 4 illustrates the user interface during transcription.

![User Interface](image.png)

Figure 4: User Interface.
The user interface consisted of an image viewer for rendering the specimen images, a form containing data fields for input with buttons for help information and submission, and a progress bar for indicating the progress/completion of tasks by the user.

The image viewer was implemented using an external JavaScript library which provided zooming and panning capabilities to the image rendered. Image could be zoomed in or out by clicking the respective buttons on the toolbar, by scrolling of the mouse or using the ‘+’ or ‘-’ keystrokes. Similarly panning or moving of the image within the viewer to the region of interest could be done by clicking the buttons on the toolbar, dragging the image using mouse or using the ‘A’ ‘D’ ‘W’ ‘S’ keystrokes to move left, right, up and down respectively. Refer to Figure 5.

Figure 5: Illustration of image zoomed and panned to specimen label.

A form containing an accordion of three collapsible panels and a button for submission of data was designed. The accordion panels were named relating to Darwin Core classes of the consisting data fields. The first accordion was labeled Taxon Information reflecting the Taxon class in Darwin Core Standard and contained two input fields Scientific Name and Scientific Name authorship (Refer to Figure 6a). The second accordion was labeled Record Information and contained two input fields Collected by and Collected Date (Refer to Figure 6b). The third accordion was labeled Location
Information and contained input fields from Darwin Core class Location. The accordion consisted of five input fields namely Location Details, Country, State Province, County/Municipality, and Locality (Refer to Figure 6c).

![Image](image1.png)

(a) Taxon Information  
(b) Record Information  
(c) Location Information

Figure 6: Illustration of user input form.

Each input fields had clickable help button on the right that opened a modal window displaying a sample label image with the corresponding data highlighted along with relevant instruction or information. In addition to the information window for individual input fields, user friendly features were added to some of the fields where applicable. Collected Date was implemented using an input box with placeholder instructing users on the required date format ("YYYY-MM-DD"). Country field was implemented using a dropdown list of Countries. Similarly, State Province input field was complemented with autocomplete feature for a set of states and provinces in Finland as the tasks were primarily from specimens collected in and around Finland. The functionality could be extended to other regions as well. Likewise, the County/Municipality
field had autocomplete features for cities and municipalities employing Google Places API [52]. Further, a general help window explaining common terms that occurred frequently on the specimen labels and their possible meaning or translations was implemented which could be loaded by clicking a button on the top left corner of the image viewer.

An indicator area was placed at the bottom of the image viewer which provided feedback for the user regarding status of the task and user’s contribution. The id of the current task along with the user’s contribution to the task using a progress bar were displayed. Refer to Figure 7.

![Figure 7: Illustration of progress bar during transcription.](image-url)

Submission of the transcription of each image label led to automatic loading of new task in the set until the user completed all the task in the set. However, users could leave transcribing of the set at any point and resume later.
5 Consensus Model

Data collected through crowdsourcing applications need to be aggregated using a consensus model to obtain a final data that can be assumed to be fit for use. The tasks were presented via the biological specimen transcription application to multiple users so as to collect redundant data for each task. The tasks when performed by multiple users could have multiple input for the same data field depending on their knowledge, experiences and interpretations. The multiple judgements for the same query from multiple users originating from the users’ own biases need to be unified through a systematic process the resultant of which is optimal if not exact.

The process of obtaining a final aggregated value for individual data fields in transcription that could be vetted against the ground data for measuring the quality of data obtained from crowdsourcing encompassed of several sub processes performed in a successive manner. The user data from crowdsourcing application stored in the database was extracted and subjected to data cleansing treatment which was then converted to the file format acceptable to the consensus building process. The individual field data entity was then algorithmically processed to produce alignment sequences. The sequences were then processed to remove outliers to generate alignment sequences from which a consent could be reached. The values obtained through such a consensus model were then evaluated against the corresponding data fields from ground data prepared in format compatible for evaluation.

The details of the sub processes involved have been discussed in the subsequent chapters with a diagrammatic overview of the entire process presented in the Figure 8.
Figure 8: Overview of processes involved in generating consensus among crowdsourced data for verification against ground data.
The data from tasks presented through the biological specimen transcription application was stored in the PyBossa database maintained in PostgreSQL server. The data from transcription for each set or individual tasks could be accessed via the API provided by the PyBossa framework. The task data after the completion of transcription tasks was extracted in JSON format. The JSON encoded data consisted of key-value pair for each data field that could be mapped into corresponding Darwin Core fields along with relevant details concerning application details, task details and user details. This data could be further processed to meet the data specification of the subsequent processes. A sample raw JSON data has been illustrated in the Figure 9.

The data extracted from database in JSON encoded format was subjected to a data cleansing treatment via a preprocessing module for removal of minor discrepancies. The module implemented in Java programming language extracted individual data fields from JSON file and performed general string manipulation operations like removal of multiple white spaces and replacement of white spaces by underscore character “_” for creation of a continuous string for each data fields. The data was then stored in tabular format mapped into corresponding Darwin Core fields into spreadsheets for improved readability during analysis and further algorithmic treatment. Refer to Figure 10 for an example image of processed data in tabular format.
The task of generating an acceptable transcription from a set of transcriptions for a particular data field is analogous to the task of identifying homologous regions in biological sequences [11, 12] and hence a cross-discipline approach to the problem was applied. The subsequent steps and the tools in the process of obtaining a consented value from a redundant set of transcriptions that can be assumed to be the result of the collective intelligence or perception of the crowd was borrowed from bioinformatics where the techniques and tools for the alignment of biological sequences have been formulated and tremendous advancements have been made. The aligning of strings is to a higher degree similar to the alignment of biological sequences in bioinformatics and hence Multiple Sequence Alignment technique was employed to obtain string alignments for transcriptions. A calculated sequence of characters from a redundant set of algorithmically aligned transcription was then produced which could be vetted with the ground data for assessment. The aforementioned problem of multiple sequence alignment, the algorithmic approach to the problem, the data standard and format used and the process of drawing consensus has been detailed in the subsequent texts.

5.1 Sequence Alignment

Sequence Alignment is the technique of arranging two or sequences of the same or different lengths to identify similar characters or regions from the sequences[53, 54]. The alignment of sequences can be done locally or globally [55]. Local alignments are helpful in identifying similar subsequences or regions of maximum similarities within the sequences. In contrast, the sequences are aligned in entirety in global alignment [56]. Global alignment methods can be customized to produce local alignments. The requirement of the current study to produce a consented sequence rather than to detect similarities within the sequences is best served by global alignments and hence the subsequent study has been focused on global alignments.

Given, a finite, not-empty set $\sum$ of elements $s_i$, which in the current context are letters

![Table](image.png)

Figure 10: Illustration of processed data in tabular format.
and symbols, a string \( S \) on \( \sum \) is a finite sequence \( S = s_1 s_2 s_3 \ldots s_n \) of elements from \( \sum \). The length of string \( S \) denoted by \( |S| \) is the number of elements in \( S \).

The terms string and sequence are usually treated synonymously in bioinformatics and shall be done so in the present text hereafter.

### 5.2 Pairwise alignment

Let \( S_1 \), \( S_2 \) be two sequences of length \( m \) and \( n \).

An alignment function maps them to sequences \( S'_1 \), \( S'_2 \) such that

\[
|S'_1| = |S'_2| = l, \text{ where } l \text{ is the length of resulting sequences}
\]

The removal of spaces from \( S'_i \) leaves \( S_i \), for \( 1 \leq i \leq 2 \).

An alignment can be obtained by insertion of spaces or gaps in any required position in the sequences such that the resulting sequences are of equal length. A gap is treated as a new character in the aligned sequence \( S'_i \). The gap in one sequence \( S'_1 \) opposite to a character \( x \) at position \( p \) in the other sequence \( S'_2 \) can be viewed as a deletion of character \( x \) at position \( p \) from the second sequence \( S'_2 \) or as an insertion of \( x \) at position \( p \) into the first sequence \( S'_1 \). An identical character \( x \) at the same position \( p \) in both the sequences is considered a match and two nonidentical characters at the same position \( p \) of the two sequences are considered a mismatch and can be treated as a replacement [57].

Given a sequence \( S_i = \text{ACDBD} \) and \( S_2 = \text{AWVB} \), an alignment \( A \) is obtained by insertion of gaps (‘−’) so that the resulting sequences \( S'_1 \) and \( S'_2 \) are of equal length and can be mapped onto each other such that each character in one sequence is opposite to a unique character in the other. A pair of sequences can be aligned in multiple ways by insertion of gaps.

For example:

Given \( S_1 = \text{ACDBD} \) and \( S_2 = \text{AWVB} \)

One possible alignment \( A_1 \) is

\[
S'_1 = \text{A C D − B D} \\
S'_2 = \text{A W − V B −}
\]
Another possible alignment $A_2$ is

\[ S'_1 = A C - D B D \]
\[ S'_2 = A W V - B - \]

Similarly, a number of alignments can be generated from which the “best” alignment can be selected. The objective of sequence alignment technique is to select the “best” alignment out of the possible alignments on the basis of a scoring scheme or function that gives a measure of similarity.

Given, $\sum$ a set of elements in the sequence $S_1$ and $S_2$, and if

\[ \sum' = \sum \cup - \]

where the character ‘−’ denotes a gap, then for any two characters $x, y$ in $\sum'$, the function $s(x, y)$ returns the score of aligning character $x$ against $y$.

A simple similarity scoring scheme can be defined as one that return a score of 1 if the characters at the same position are identical, -1 if the characters are not identical and -2 if a gap ‘−’ is encountered at the position. In a typical biosequence alignments, the similarity scoring functions are substitution matrices with similarity measures of each character-pair [57]. There are numerous such substitution matrices suggested in bioinformatics [58], however for the sake of simplicity, we have considered the aforementioned simple scoring scheme presented in tabular format. Refer to Table 5.

Table 5: A simple scoring scheme

<table>
<thead>
<tr>
<th>Condition</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match</td>
<td>+1</td>
</tr>
<tr>
<td>Mismatch</td>
<td>-1</td>
</tr>
<tr>
<td>Gap</td>
<td>-2</td>
</tr>
</tbody>
</table>

For a given alignment $A$ of sequences $S_1$ and $S_2$, if $S'_1$ and $S'_2$ are the sequences formed after insertion of gaps into $S_1$ and $S_2$ respectively where

\[ |S'_1| = |S'_2| = l, \]
then, the value of the alignment $A$ is defined as

$$\sum_{i=1}^{l} s(S'_1(i), S'_2(i))$$

Here, every position $i$ in alignment $A$ specifies a pair of characters from $\sum'$ and the value of $A$ is obtained from the summation of the character-pair value of each pair. In alignment governed by such a scoring scheme, the matches between the two strings are emphasized while the mismatches or gaps are penalized. The penalty for a gap of length $g$ is defined either by a linear score

$$\gamma(g) = -gd$$

or an affine score

$$\gamma(g) = -d - (g - 1)e$$

where $d$ is the gap open penalty and $e$ is the gap extension penalty.

The selection of alignment is performed on the basis of higher alignment value. For a predefined scoring function, the similarity of the two sequences $S_1$ and $S_2$ is defined as the value of the alignment $A$ of $S'_1$ and $S'_2$ which maximizes the total alignment value. It is also referred to as the optimal alignment value of $S_1$ and $S_2$.

It is obvious that a brute force approach to the task of selecting the optimal alignment is dependent on the number of possible alignments. A brute force approach of enumerating all possible alignments, scoring them and selecting the alignment with highest score is not practically feasible except for sequences of small length as the number of possible alignments grows rapidly with the length of the sequences [59, 60].

Given sequences $S_1$ and $S_2$ of length $n$ and $m$, the number of possible alignments is given by

$$N(n, m) = \sum_{r=0}^{\min(n,m)} \binom{n}{r} \binom{m}{r} = \binom{n+m}{r}$$

For $n = m$

$$N(n, n) = \sum_{r=0}^{n} \binom{n}{r} \binom{n}{r} = \binom{2n}{n}$$
which can be approximated by \( \frac{2^n}{\sqrt{\pi n}} \) using Stirling’s formula.

This demonstrates that the number of alignments grows at least exponentially with the length of the sequences [60, 59]. The exponential growth of the possible alignment requires efficient methods for selection of optimal alignment and hence the solution to the problem can be sought using dynamic programming approach [61, 62, 53, 63].

5.2.1 The Needleman-Wunsch algorithm

Needleman-Wunsch algorithm is a dynamic programming solution to pairwise alignment which reduces the number of possible alignments to be evaluated and yet guarantees the best result or the optimal alignment [62]. It is based on the fundamental idea of generating the best alignment using the optimal alignments of smaller subsequences.

Given two sequences \( S_1 = (x_1, x_2, \ldots, x_n) \) and \( S_2 = (y_1, y_2, \ldots, y_m) \),

A scoring matrix \( F \) is computed.

\[
F : \{1, 2, \ldots, n\} \times \{1, 2, \ldots, m\} \rightarrow \mathbb{R}
\] (1)

where, \( F(i; j) \) equals the optimal value of the alignment of two prefixes \((x_1, x_2, \ldots, x_i)\) and \((y_1, y_2, \ldots, y_j)\).

This is done recursively by computing \( F(i, j) \) from \( F(i - 1, j - 1) \), \( F(i - 1, j) \) and \( F(i, j - 1) \). The final value \( F(n, m) \) contains the score of the optimal alignment between \( S_1 \) and \( S_2 \). The alignment corresponding to this score is obtained through back tracing the path of choices that the recursion made.

The algorithm consists of three steps:

The first step is the creation of score matrix and initialization of the first row and column such that the scores account for gap penalties as defined by

\[
\begin{align*}
F(0, 0) &= 0 \\
F(i, 0) &= F(i - 1, 0) - d \\
F(0, j) &= F(0, j - 1) - d
\end{align*}
\]

The second step is the recursive computation of scores and population of the score and traceback matrix by traversing through the entries \((i, j)\) of scoring matrix determining
the optimal alignment of the sequences $S_1[1, \ldots, i]$ and $S_2[1, \ldots, j]$ using the entries $(i - 1, j - 1)$, $(i - 1, j)$, and $(i, j - 1)$. The pointer to the corresponding entry from which the optimal alignment at $(i, j)$ was built is added to the traceback matrix.

$$ F(i, j) = \max \begin{cases} 
F(i - 1, j) - g & \text{align } x_i \text{ with a gap} \\
F(i, j - 1) - g & \text{align } y_i \text{ with a gap} \\
F(i - 1, j - 1) + s(x_i, y_j) & \text{align } x_i \text{ with } y_i \text{ match or mismatch}
\end{cases} \tag{3} $$

where,

$s(x_i, y_j)$ is the similarity or substitution score for residues $i$ and $j$, and $g$ is the gap penalty.

The final step is to construct the optimal alignment by following the path of pointers starting at entry $(n, m)$ of the traceback matrix and ending at entry $(0, 0)$ as the score of the optimal alignment is contained in entry $(n, m)$ of the scoring matrix [62].

It is important to consider the time and space complexity of the algorithm. Regarding the complexity of the algorithm, there are $n \times m$ elements in the matrix $F$, and each element require a constant number of calculations to compute and hence the total running time is $O(nm)$. Similarly, the total space complexity is $O(nm)$. For $n = m$ the running time is $O(n^2)$ [62].

The pseudocode representation of the algorithm is presented in Algorithm 1.
Algorithm 1 Needleman-Wunsch algorithm

Input: Two sequences $S_1$ and $S_2$

Output: Optimal alignment and score of $S_1$ and $S_2$

Initialization:
Set $F(i, 0) := -i.g$ for all $i = 0, 1, 2, \cdots, n$
Set $F(0, j) := -j.g$ for all $j = 0, 1, 2, \cdots, m$

for $i = 1, 2, \cdots, n$ do
  for $j = 1, 2, \cdots, m$ do
    Set $F(i, j) := \max \left\{ \begin{array}{l} F(i-1, j) - g \\ F(i, j-1) - g \\ F(i-1, j-1) + s(x_i, y_j) \end{array} \right.$
    Set corresponding backtrace matrix $T(i, j)$ to the maximizing pair $(i'; j')$.
  end for
end for

The optimal score $\alpha := F(n, m)$
Set $(i, j) := (n, m)$

while $(i > 0 \text{ or } j > 0)$ do
  if $T(i, j) = (i - 1; j - 1)$ then
    return $(x_i, y_j)$
  else if $T(i, j) = (i - 1, j)$ then
    return $(x_i, \_)$
  else if $T(i, j) = (i, j - 1)$ then
    return $(\_, y_j)$
  end if
  Set $(i, j) := T(i, j)$
end while
5.3 Multiple Sequence Alignment

The concept of sequence alignment can be extended to more than two sequences. Given a set of sequences \( S = S_1, S_2, \ldots, S_k \) a global multiple alignment \( A \) is produced by inserting gaps in each of the \( k \) sequences making them of equal length \( l \) so that the sequences can be arranged into \( k \) rows and \( l \) columns each such that each character of each of the sequences is in a unique column.

One of the most commonly used approach to the task is by computing an alignment that maximizes the sum of similarities for all pairs of sequences known as the sum-of-pairs score \([54, 57]\). The sum-of-pairs (SP) value for a multiple global alignment \( A \) of \( k \) sequences is the sum of the values of all \( \binom{k}{2} \) pairwise alignments induced by \( A \). An optimal SP (global) alignment of sequences \( S_1, S_2, \ldots, S_k \) is an alignment that has the optimal sum-of-pairs value for these sequences \([54]\).

The complexity of this approach using sum-of-pairs scales exponentially with the number of sequences \([64]\). For sequences \( k \) of length \( n \), it would require a \( k \)-dimensional matrix of size \((n + 1)^k\). A \( k \)-dimensional cube has \(2^k\) vertices and hence the maximum of \(2^k - 1\) neighboring cells for each entry in the score matrix has to be computed. This results in the complexity of computing the optimal alignment of \( O((2n)^k) \) which is exponential in the order of the number of sequences to be aligned \([65, 66]\).

In addition to this, the optimal sum-of-pair alignment problem has been shown to be NP-Complete \([57, 67]\). The complexity in exponential order of the number of sequences and the fact that the problem is NP-Complete, a number of heuristic methods \([68, 66, 69, 65, 70, 71, 72, 69]\) with approximation approach have been devised over the years. The current study employed MAFFT, a multiple sequence alignment method based on the Fast Fourier Transform (FFT). The input data format used in this current study for sequence alignment, the FASTA format and the MAFFT program has been presented in the subsequent chapters.
5.3.1 FASTA format

FASTA format also known as Pearson format is a text-based format for representing either nucleotide sequences or peptide sequences in bioinformatics studies [73, 74]. The base pairs or amino acids are represented using single-letter codes in the format. The format owes its origin to the protein sequence alignment software package FASTP [73].

A sequence in FASTA format consists of a single-line description, a carriage return, followed by any number of lines of sequence data [74]. A greater-than (">") symbol at the beginning marks the description line distinguishing it from the sequence data. Conventionally, the description line or the header line assigns a name and/or a unique identifier to the following sequence and in some cases contain additional information about the sequence. The sequence data consist of one or more lines of codes following the header with each line recommended to limit to 80 characters. The sequence can contain gaps or alignment characters with a single hyphen or dash used for the representation of gaps [75].

A FASTA file can contain more than one sequence with multiple sequences in FASTA format concatenated in a list [74]. However, the format prohibits blank lines in the middle of FASTA input. A sample FASTA formatted sequences of string is illustrated in the Figure 11.

```
>1-6-8-1-6
Kn._Sotkamo._Ontojoki._Romuvaaran_paallys,_keskustassa.
>1-6-8-7-186
Kn._Sotkamo._Ontoriver
>1-6-8-3-226
Kn._Sotkamo_Ontojoki_Romuvaaran_Paallys_Keskustassa
>1-6-8-19-251
Kn._Sotkamo._Ontojoki,_Romuvaaran_paallys,_Keskustassa
>1-6-8-21-279
Romuvaaran_paallys,_keskustassa.
```

Figure 11: Illustration of data in FASTA format.
Multiple Alignment using Fast Fourier Transform (MAFFT)

Multiple Alignment using Fast Fourier Transform (MAFFT) is a multiple sequence alignment program based on Fast Fourier Transformation [76, 77, 78, 79, 68]. It consists of a unique technique of detecting homologous sections by Fast Fourier Transform and a similarity scoring system that is both efficient and accurate even for sequences with large gaps. It was primarily designed using a fast group-to-group alignment algorithm based on Fast Fourier Transform (FFT) and the 6mer method, an approximate distance calculation method [68]. The program has evolved implementing objective functions like the weighted sum of pairs score or a consistency score calculated from all-to-all pairwise alignments improving the accuracy along with efficiency. The program provides a number of options using different strategies to address diverse multiple sequence alignment problems [76].

The program best suited for the current study was the Iterative refinement method referred to as the FFT-NS-i option in MAFFT. The steps involved in the method is illustrated in the Figure 12.

![Figure 12: Steps in MAFFT FFT-NS-i. [80](image)](image)

The first step consists of generating a distance matrix based on the number of k-tuples shared by the sequences. The distance matrix is then used to generate a guide tree using a hierarchical clustering method. The next step is aligning the sequences in a progressive manner in order of the branching of sequences in the guide tree. This is
achieved using the FFT algorithm and the similarity matrix. This process is referred to as FFT-NS-1. In the following step, a second stage guide tree is constructed using the alignment from FFT-NS-1. This guide tree so produced is used to generate alignment from the input sequences which can be expected to more reliable. This technique is referred to FFT-NS-2. The next step in the process is referred to as FFT-NS-i. It is an iterative process in which the alignment obtained from the FFT-NS-2 is divided into two groups using tree dependent restricted partitioning and then realigned. This process of division and realignment is repeated to the point when the score cannot be further improved [77].

The bash script implemented for MAFFT program has been detailed as follows.

```bash
#!/bin/bash
if [[ "$#" -ne 2 ]]; then
    exit 1
fi
inputFolder="$1"
outputFolder="$2"
for currentfile in "$inputFolder"/*
    do
        onlyFilename=$(basename "$currentfile")
        fftsnsi --text $currentfile > $outputFolder/$onlyFilename
    done
```

A sample alignment generated by the MAFFT FFT-NS-i in FASTA format is illustrated in Figure 13. Please note the gaps inserted in the aligned sequences.

```
>1-6-8-1-6
Kn._Sotkamo._Ontojoki._Romuvaaran_paallys,_keskustassa.
>1-6-8-7-186
Kn._Sotkamo._Ontoriver-------------------------------
>1-6-8-3-226
Kn._Sotkamo._Ontojoki._Romuvaaran_Paallys._Keskustassa-
>1-6-8-19-251
Kn._Sotkamo._Ontojoki._Romuvaaran_paallys,_Keskustassa-
>1-6-8-21-279
---------------------Romuvaaran_paallys,_keskustassa.
```

Figure 13: Illustration of MAFFT FFT-NS-i aligned data.
5.4 Outliers Removal

The next step in building the consensus sequence was the post processing of the alignments for outlier removal. During the visual inspection of the alignments, it was observed that random or singular characters appeared within the alignments which were the result of garbage entries from users. Such rogue entries could be treated as outliers [11] and hence a simple method for the removal of such singular occurrence from the alignment was devised.

The singular entries were detected in the alignment by scanning through the sequences in the alignment each position at a time. If a sequence contained character corresponding to which all the other sequences in the alignment contained gaps in the position, the character at that particular position in the sequence would be treated as an outlier and replaced by a gap. This process helped to improve the consensus among the sequences.

The code snippet of the procedure implemented in Java programming language has been detailed as follows.

```java
ArrayList<String> idList = new ArrayList<String>();
ArrayList<String> contentList = new ArrayList<String>();
FastaSequence currentFASTA = new FastaSequence();
currentFASTA.readSequenceFromFile(currentFile.getAbsolutePath());
idList = currentFASTA.getDescription();
contentList = currentFASTA.getSequence();
for (int lineIter = 0; lineIter < contentList.size(); lineIter++) {
    String currentLine = contentList.get(lineIter);
    for (int charIter = 0; charIter < currentLine.length(); charIter++) {
        String currentChar = String.valueOf(currentLine.charAt(charIter));
        if (!"-".equals(currentChar) && 
            ".equals(currentChar)) {
            boolean replaceOrNot = false;
            for (int k = 0; k < contentList.size(); k++) {
                if (k != lineIter) {
                    String compChar = String.valueOf(contentList.get(k).charAt(charIter));
                    if (currentChar.equals(compChar)) {
                        if (compChar != ")" || compChar != "])" || compChar != "]") { // Replace the outlier
                            currentChar = ");
                        }
                    }
                }
            }
        }
    }
}
```
if (!"-".equals(compChar) || 
  "_".equals(compChar))) {
  replaceOrNot = true; break;
}

if (replaceOrNot == false) {
  if (!"-".equals(currentChar) || 
      "_".equals(currentChar))) {
    String temp =
      replaceCharAt(contentList.get(lineIter),
      charIter, "-");
    contentList.remove(lineIter);
    contentList.add(lineIter, temp);
    replaceCharAt(contentList.get(lineIter),
      charIter, "-");
  }
}

5.5 Consensus

The final task in the processing of the crowdsourced data was to generate consensus from the alignment sequences. The consented value from the alignments corresponded to individual data fields from the redundantly presented tasks in the crowdsourcing platform. The consented values from the alignments were achieved using a consensus building function dumb consensus implemented in Python [81, 82, 11].

The function has been implemented such that each sequence was read throughout its length one character position at a time counting the number of occurrence of unique characters at the same position in the alignment. If the percentage of the most frequently occurring character at the position exceeded a specific threshold, the character was accepted and added to the final consensus sequence. In case the specified threshold could not be exceeded by any character at the position in the sequences, then a char-
acter indicating ambiguity “X” was added to the consensus string [82]. The threshold percentage for acceptance was set to 50 in the current study.

The python script implemented for the task with necessary arguments has been detailed as follows.

```python
statinfo = os.stat(current_file)
if (statinfo.st_size > 0):
    alignment = AlignIO.read(open(current_file), "fasta",
        alphabet = Alphabet.generic_alphabet)
    align_info = AlignInfo.SummaryInfo(alignment)
    consensus = align_info.dumb_consensus(0.5)
    with open(consensusfile, "w") as f:
        f.write(consensus._data)
else:
    with open(consensusfile, "w") as f:
        f.write(""")
```
6 Results

The data collected from the crowdsourcing application on being processed through the workflow detailed in Figure.8 produced consented values that were evaluated against the ground data. The evaluation of similarity between the ground data and the data from the consensus model was performed by measuring the Levenshtein distance between the two.

The similarity between the strings from expert-transcribed data and consented value from the crowdsourced data was measured using the widely used string metric, the Levenshtein distance [83, 84]. Given two strings, the Levenshtein distance between the two is the minimum number of single-character changes, the changes being insertion, deletion or substitution of a character at a position, that is required to transform one string into the other.

For every pair of string \( S_1 \) and \( S_2 \) from expert data and consensus data respectively, and the Levenshtein distance \( \text{lev}(S_1, S_2) \) was calculated as defined as follows.

\[
\text{lev}(S_1, S_2)(i, j) = \begin{cases} 
\max(i, j) & \text{if } \min(i, j) = 0 \\
\min \begin{cases} 
\text{lev}(S_1, S_2)(i - 1, j) + 1 \\
\text{lev}(S_1, S_2)(i, j - 1) + 1 \\
\text{lev}(S_1, S_2)(i - 1, j - 1) + 1(a_i \neq b_j)
\end{cases} & \text{else}
\end{cases}
\]  

(4)

Further, the similarity ratio between the two strings was calculated as follows.

\[
\text{sim}(S_1, S_2) = 1 - \frac{\text{lev}(S_1, S_2)}{\max(|S_1|, |S_2|)}
\]  

(5)

The data used as the ground truth in the study was data generated from transcription by biological experts at Digitarium, University of Eastern Finland. The tasks presented to the participants of the crowdsourcing application had been prior transcribed and verified by experts at Digitarium, and hence the performance of the crowd could be assessed against the experts.

The data used as the ground truth were stored in XML files with each file representing a unique specimen presented in the task. The data was maintained in the XML files in Darwin Core format. The ground data from the individual XML files were extracted
and populated in a spreadsheet. The spreadsheets storing the consensus data and the ground data were made compatible for ease in visualization and processing. Figure 14 illustrates the ground data in standard Darwin Core format.

![Figure 14: Illustration of ground data in Darwin Core XML format.](image)

The mean, median and the standard deviation of the similarity ratio of each of the data fields was calculated as illustrated in Figure 15.
The mean and the median values complemented with the standard deviation of similarity ratio of individuals field presented in Figure 15 can lead us to conclusions regarding the accuracy of the transcription at individual datafield level. It can be observed from the figure that the accuracy of transcription in some of the fields namely scientificName, recordedBy, eventDate, country and stateProvince were acceptable where as the accuracy in others viz. scientificNameAuthorship, verbatimLocality, county and locality were below par. The results at individual level have been discussed in the subsequent texts.

The mean and median value of the similarity ratio of scientificName field is on the higher end along with a lower standard deviation. This indicated higher accuracy in the transcription of scientificName. This could be due to the fact that the values for scientificName are easily identifiable from the label as they are normally Latin words which appear out in the label and the collectors emphasize on them while recording as well.

In contrast, the accuracy in the transcription of scientificNameAuthorship field was moderate. The standard deviation and the median value of the field indicate the average performance of the participants. The plausible reason for this tendency could be the fact that the scientificNameAuthorship occurs normally at the end of scientificName in the specimen labels and are in abbreviated forms in a number of instances which makes them indistinguishable from scientificName in some cases.
The accuracy was high in the transcription of `recordedBy` field as illustrated in the plot. The possible explanation for this is the fact that the collector’s name appear identifiable as human names have distinctive characters in comparison to other terms that normally occur in the labels. Further, the collector’s name is normally preceded by the term “leg.” which makes the distinction easy.

Likewise, the `eventData` field was also transcribed with a high degree of accuracy. The fact that the data field is distinctive in the label image due to its uniqueness and the adherence to instruction to enter the data in a specific format by the participants is reflected by the level of accuracy in transcription.

The `verbatimLocality` field was the most difficult for users during transcription and is demonstrated by the unacceptably low accuracy of the user’s transcription. The field requires a higher level of understanding and interpretation as most amateur participant fail to distinguish between the locality related information and habitat information present on the label image. In addition to this, the fact that some image labels were documented in languages beyond participant’s understanding contributed in creating the confusion during transcription.

On the contrary, the accuracy in `country` field was toward the maximum as can be observed from the plot. Despite the common trend that the country where the specimen was collected was not mentioned in the labels by the collector, participants were successful in interpreting it from the existing locality information on the label. The drop-down feature implemented for selecting the country rather than typing it in the user interface is the possible reason for such high level of accuracy in transcription of the field.

Similarly, the accuracy of transcription of the `stateProvince` field, as can be seen from the plot is high. The implementation of the autocomplete feature for the data field as the number of possible values for the geographical unit is limited, appear to be the reason behind the achieved accuracy.

The accuracy in the transcription of the `county` field for transcription of geographical units at country and municipalities level is low. Despite the implementation of a suggestive autocomplete feature based on Google Places API, the accuracy of transcription is low. The variation in the interpretation of the geographical unit among participants could be the reason for the lower accuracy despite the augmentation.
Further, the accuracy of transcription of another geographical unit, the *locality* field is low in a similar manner. The low accuracy in transcription can be explained on the basis of the fact that this field required knowledge of the local geography and interpretation of some level. This posed a bigger challenge in transcription which is demonstrated by the lower accuracy level as can be seen from the plot.

The results of the transcription and the consensus model in the form of similarity ratio demonstrated a higher level of accuracy in some of the data fields, where as in some fields, the accuracy was moderate and in some were extremely below the acceptable level.
7 Conclusion and Recommendations

The two-fold purpose of the study of implementing a crowdsourcing application for transcription of biological specimen and assessing the accuracy of data transcribed through it was successfully conducted. The application was designed and implemented using PyBossa framework and was accessible via “http://manash.eu/” for participation. The 100 tasks of transcribing biological specimens were completed by the participants and there were no major complaints from the users regarding the functionality or usability of the application.

The data collected through the application was further processed through the workflow detailed in the preceding chapter to build consensus among the redundant entries. The consensus data generated was then evaluated against the ground data for measuring the accuracy of the transcription and the consensus model described in the earlier chapters. The result of the experiment was presented and the probable reason for the tendencies observed suggested.

The result demonstrated a high degree of accuracy in some of the data fields whereas the accuracy was extremely low in some. This can lead us to the conclusion that the participants were comfortable in identifying and transcribing some data fields namely scientificName, recordedBy, eventDate, country and stateProvince whereas the data fields scientificNameAuthorship, verbatimLocality, county and locality were perceived difficult as demonstrated by their lower accuracy.

An application designed and implemented well to any degree always leaves room for improvement and the application implemented in the current study is no exception. It can be inferred from the results that features that aided the participants in transcription like the suggestive autocomplete feature for input fields and instructions to adhere to a fixed data format improved the accuracy of transcription. It is recommended to implement such features in the user interface where possible for enhancement of user data quality.

Similarly, use of controlled vocabulary namely the list of countries in the data field country field improved the accuracy of transcription. It is recommended to implement such feature where possible as it removes the possibility of error of omission or mistyping.
Like any other scientific study, the procedures described earlier have areas for further enhancements and continued development. The preprocessing and post-processing involved in the consensus building steps can be improved for better results. Likewise, the Multiple Sequence Alignment algorithm used in the current study can be investigated for parameterizing or fine tuning to meet the applied context. Different approximate multiple sequence alignment algorithms [66, 69, 65, 70, 71, 72, 69] can be explored and comparative evaluation of the accuracy can be conducted in addition to investigation of the different accuracy measures for sequences which can be subjects for further studies.
Bibliography


A generic schema of Darwin Core terms

<?xml version="1.0" encoding="UTF-8"?>
<xs:schema version="2013-10-22"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://rs.tdwg.org/dwc/terms/"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dcelem="http://purl.org/dc/elements/1.1/"
    xmlns:dwctype="http://rs.tdwg.org/dwc/dwctype/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/">
    <xs:annotation>
        <xs:documentation xml:lang="en">
            The Generic Darwin Core schema defining all property terms as global elements.
            Domain classes are not defined here, but in a separate schema (tdwg_dwcclasses.xsd) with the same namespace.
            There are two ways in which references to domains can be defined:
            a) through an abstract base term anyXXXTerm, which is derived from the type dwc:anyProperty and which all properties for that domain use as their substitution group. If you refer to dwc:anyXXXTerm in your schema, you will be able to reference any of the terms, but it will be impossible to create a sequence of all terms occurring only once.
            b) through a group of elements called XXXTerms, which reference properties exactly once as a sequence. This group can be used to refer to all the domain properties only once. In order to make a property required, you must create your own group referring to the individual dwc terms. If you want to specify every term once at most and in any order, you will also have to refer to the individual dwc term yourself — a limitation of the xs:all model.
        </xs:documentation>
    </xs:annotation>
    <!−−(tdwg_basetypes.xsd defines custom data types such as dateTimeISO to allow dates or the forms YYYY and YYYY-MM. −−)>
    <xs:include schemaLocation="tdwg_basetypes.xsd"/>
    <xs:import namespace="http://purl.org/dc/terms/"
        schemaLocation="http://dublincore.org/schemas/xmls/2008/02/11/dcterms.xsd"/>
    <xs:import namespace="http://purl.org/dc/elements/1.1/"
        schemaLocation="http://dublincore.org/schemas/xmls/dc.xsd"/>
        schemaLocation="http://www.w3.org/2001/03/xml.xsd"/>
    <xs:import namespace="http://rs.tdwg.org/dwc/dwctype/"
        schemaLocation="tdwg_dwctypes.xsd"/>
    <!−− ANY DWC TERM −→>
    <xs:element name="any" abstract="true"/>
    <!−− ANY DWC IDENTIFIER −→>
    <xs:element name="anyIdentifier" abstract="true" substitutionGroup="dwc:any"/>
    <xs:group name="IdentifierTerms">
        <xs:annotation>
            <xs:documentation>Group of all Darwin Core identifier terms as a convenience to construct other schemas</xs:documentation>
        </xs:annotation>
        <xs:sequence>
            <xs:element ref="dwc:occurrenceID" minOccurs="0"/>
        </xs:sequence>
    </xs:group>
</xs:schema>
<xs:element name="anyProperty" abstract="true" substitutionGroup="dwc:any" />
<xs:element name="anyClass" abstract="true" substitutionGroup="dwc:any" />

Every domain has its own substitutionGroup

<xs:element name="institutionID" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="collectionID" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="datasetID" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="institutionCode" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="collectionCode" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="datasetName" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="ownerInstitutionCode" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="basisOfRecord" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="informationWithheld" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="dataGeneralizations" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:element name="dynamicProperties" substitutionGroup="dwc:anyRecordLevelTerm" type="xs:string" />
<xs:group name="RecordLevelTerms">
<xs:sequence>
  <xs:element ref="dcterms:type" minOccurs="0" />
  <xs:element ref="dcterms:modified" minOccurs="0" />
  <xs:element ref="dcterms:language" minOccurs="0" />
  <xs:element ref="dcterms:rights" minOccurs="0" />
  <xs:element ref="dcterms:rightsHolder" minOccurs="0" />
  <xs:element ref="dcterms:accessRights" minOccurs="0" />
  <xs:element ref="dcterms:bibliographicCitation" minOccurs="0" />
  <xs:element ref="dcterms:references" minOccurs="0" />
  <xs:element ref="dwc:institutionID" minOccurs="0" />
  <xs:element ref="dwc:collectionID" minOccurs="0" />
  <xs:element ref="dwc:datasetID" minOccurs="0" />
  <xs:element ref="dwc:institutionCode" minOccurs="0" />
  <xs:element ref="dwc:collectionCode" minOccurs="0" />
  <xs:element ref="dwc:datasetName" minOccurs="0" />
</xs:sequence>
<xs:element ref="dwc:associatedMedia" minOccurs="0"/>
<xs:element ref="dwc:associatedReferences" minOccurs="0"/>
<xs:element ref="dwc:associatedOccurrences" minOccurs="0"/>
<xs:element ref="dwc:associatedSequences" minOccurs="0"/>
<xs:element ref="dwc:associatedTaxa" minOccurs="0"/>
</xs:sequence>
</xs:group>

<!-- MaterialSample domain -->
<xs:element name="anyMaterialSampleTerm" abstract="true" substitutionGroup="dwc:anyProperty"/>
<xs:element name="materialSampleID" type="dwc:nonEmptyString" substitutionGroup="dwc:anyIdentifier"/>

<!-- EVENT domain -->
<xs:element name="anyEventTerm" abstract="true" substitutionGroup="dwc:anyProperty"/>
<xs:element name="eventID" type="dwc:nonEmptyString" substitutionGroup="dwc:anyIdentifier"/>

<!-- domain properties -->
<xs:element name="samplingProtocol" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="samplingEffort" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="eventDate" type="dwc:dateTimeISO" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="eventTime" type="xs:time" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="startDayOfYear" type="dwc:dayOfYearDataType" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="endDayOfYear" type="dwc:dayOfYearDataType" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="year" type="xs:gYear" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="month" type="xs:gMonth" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="day" type="xs:gDay" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="verbatimEventDate" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="habitat" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="fieldNumber" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="fieldNotes" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
<xs:element name="eventRemarks" type="xs:string" substitutionGroup="dwc:anyEventTerm"/>
</xs:group>

<!-- dcterms:LOCATION domain -->
<xs:element name="anyLocationTerm" abstract="true" substitutionGroup="dwc:anyProperty"/>
<xs:element name="locationID" type="dwc:nonEmptyString" substitutionGroup="dwc:anyIdentifier"/>

<!-- domain properties -->
<xs:element ref="dwc:samplingProtocol" minOccurs="0"/>
<xs:element ref="dwc:samplingEffort" minOccurs="0"/>
<xs:element ref="dwc:eventDate" minOccurs="0"/>
<xs:element ref="dwc:eventTime" minOccurs="0"/>
<xs:element ref="dwc:startDayOfYear" minOccurs="0"/>
<xs:element ref="dwc:endDayOfYear" minOccurs="0"/>
<xs:element ref="dwc:year" minOccurs="0"/>
<xs:element ref="dwc:month" minOccurs="0"/>
<xs:element ref="dwc:day" minOccurs="0"/>
<xs:element ref="dwc:verbatimEventDate" minOccurs="0"/>
<xs:element ref="dwc:habitat" minOccurs="0"/>
<xs:element ref="dwc:fieldNumber" minOccurs="0"/>
<xs:element ref="dwc:fieldNotes" minOccurs="0"/>
<xs:element ref="dwc:eventRemarks" minOccurs="0"/>
</xs:sequence>
</xs:group>

<!-- dcterms:LOCATION domain -->
<xs:element name="anyLocationTerm" abstract="true" substitutionGroup="dwc:anyProperty"/>
<xs:element name="locationID" type="dwc:nonEmptyString" substitutionGroup="dwc:anyIdentifier"/>

<!-- domain properties -->
<xs:element name="higherGeographyID" type="dwc:nonEmptyString" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="higherGeography" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="continent" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="waterbody" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="islandGroup" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="island" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="country" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="countryCode" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="stateProvince" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="county" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="locality" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimLocality" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimElevation" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="minimumElevationInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="maximumElevationInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimDepth" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="minimumDepthInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="maximumDepthInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="minimumDistanceAboveSurfaceInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="maximumDistanceAboveSurfaceInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="locationAccordingTo" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="locationRemarks" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimCoordinates" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimLatitude" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimLongitude" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="verbatimCoordinateSystem" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="decimalLatitude" type="dwc:decimalLatitudeDataType" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="decimalLongitude" type="dwc:decimalLongitudeDataType" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="geodeticDatum" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="coordinateUncertaintyInMeters" type="xs:double" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="coordinatePrecision" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="pointRadiusSpatialFit" type="dwc:spatialFitDataType" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="footprintWKT" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="footprintSRS" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="footprintSpatialFit" type="dwc:spatialFitDataType" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="georeferencedBy" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="georeferencedDate" type="dwc:dateTimeISO" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="georeferenceProtocol" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="georeferenceSources" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="georeferenceVerificationStatus" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:element name="georeferenceRemarks" type="xs:string" substitutionGroup="dwc:anyLocationTerm"/>
<xs:group name="LocationTerms">
  <xs:sequence>
    <xs:element ref="dwc:higherGeographyID" minOccurs="0"/>
    <xs:element ref="dwc:higherGeography" minOccurs="0"/>
    <xs:element ref="dwc:continent" minOccurs="0"/>
  </xs:sequence>
</xs:group>
<xs:element ref="dwc:relationshipRemarks" minOccurs="0"/>
</xs:sequence>
</xs:group>

<!−− MEASUREMENTORFACT domain −−>
<xs:element name="anyMeasurementTerm" abstract="true" substitutionGroup="dwc:anyProperty"/>
<xs:element name="measurementID" type="dwc:nonEmptyString" substitutionGroup="dwc:anyIdentifier"/>

<!−− domain properties −−>
<xs:element name="measurementType" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementValue" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementAccuracy" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementUnit" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementDeterminedDate" type="dwc:dateTimeISO" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementDeterminedBy" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementMethod" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:element name="measurementRemarks" type="xs:string" substitutionGroup="dwc:anyMeasurementTerm"/>
<xs:group name="MeasurementTerms">
<xs:sequence>
  <xs:element ref="dwc:measurementType" minOccurs="0"/>
  <xs:element ref="dwc:measurementValue" minOccurs="0"/>
  <xs:element ref="dwc:measurementAccuracy" minOccurs="0"/>
  <xs:element ref="dwc:measurementUnit" minOccurs="0"/>
  <xs:element ref="dwc:measurementDeterminedDate" minOccurs="0"/>
  <xs:element ref="dwc:measurementDeterminedBy" minOccurs="0"/>
  <xs:element ref="dwc:measurementMethod" minOccurs="0"/>
  <xs:element ref="dwc:measurementRemarks" minOccurs="0"/>
</xs:sequence>
</xs:group>
</xs:schema>
B Darwin Core classes

<?xml version="1.0" encoding="UTF-8"?>
<xs:schema version="2013-10-22" xmlns:xs="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://rs.tdwg.org/dwc/terms/"
    xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
    xmlns:include="schemaLocation="tdwg_dwcterms.xsd"/>

<xs:annotation>
    <xs:documentation>
        This schema redefines all darwin core classes to disallow compound
        objects — within a class no other nested class is allowed. It
        allows properties to occur in any order and as many times as needed.
    </xs:documentation>
</xs:annotation>

<!-- OCCURRENCE domain -->
<xs:element name="Occurrence" substitutionGroup="dwc:anyClass">
    <xs:complexType>
        <xs:choice maxOccurs="unbounded">
            <xs:element ref="dwc:anyIdentifier"/>
            <xs:element ref="dwc:anyOccurrenceTerm"/>
            <xs:element ref="dwc:anyRecordLevelTerm"/>
        </xs:choice>
    </xs:complexType>
</xs:element>

<!-- MATERIALSAMPLE domain -->
<xs:element name="MaterialSample" substitutionGroup="dwc:anyClass">
    <xs:complexType>
        <xs:choice maxOccurs="unbounded">
            <xs:element ref="dwc:anyIdentifier"/>
            <xs:element ref="dwc:anyMaterialSampleTerm"/>
            <xs:element ref="dwc:anyRecordLevelTerm"/>
        </xs:choice>
    </xs:complexType>
</xs:element>

<!-- EVENT domain -->
<xs:element name="Event" substitutionGroup="dwc:anyClass">
    <xs:complexType>
        <xs:choice maxOccurs="unbounded">
            <xs:element ref="dwc:anyIdentifier"/>
            <xs:element ref="dwc:anyEventTerm"/>
            <xs:element ref="dwc:anyRecordLevelTerm"/>
        </xs:choice>
    </xs:complexType>
</xs:element>

<!-- LOCATION domain -->
<!-- reuses the dc:location term, so see dublin_core.xsd for definition -->
<!-- -->
<xs:element name="Location" substitutionGroup="dwc:anyClass"/>
 <! -- MEASUREMENTORFACT domain -- >
<xs:element name="MeasurementOrFact" substitutionGroup="dwc:anyClass">
    <xs:complexType>
        <xs:choice maxOccurs="unbounded">
            <xs:element ref="dwc:anyIdentifier"/>
            <xs:element ref="dwc:anyMeasurementTerm"/>
            <xs:element ref="dwc:anyRecordLevelTerm"/>
        </xs:choice>
    </xs:complexType>
</xs:element>
</xs:schema>