Never let school interfere with your education.
Mark Twain

To succeed in life, you need three things: a wishbone, a backbone, and a funnybone.
Reba McEntire

Teardrops dripping down, an eternal rivers flow.
Distress concealed behind a smile, the world doesn't need to know.
A swamp full of sorrows, for miles and miles on end.
Another locked in eternal sleep, to join the earth again.
Cindy M. Wong

University of Alberta

Human-Based Computation for Microfossil Identification

by

Cindy M. Wong

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

Department of Electrical and Computer Engineering

Edmonton, Alberta
Fall 2011

To all of my grandparents, both here and gone. To my parents, my family, for all that you've done. Your love and support have meant the world, so this is to you from your little girl.

Abstract

Image understanding is a general challenge in Artificial Intelligence (AI) because of its complexity. It is considered an AI-complete problem. We focus on the specific, important, and difficult case of microfossil identification, which is currently done manually. Microfossil identification has applications in paleoenvironmental research and oil exploration. We use evolutionary prototyping to engineer a complex system that employs crowdsourcing, mainly human-based computation. Our latest prototype, called the Microfossil Quest, combines human intelligence, including expert and citizen science, with computer intelligence, including unsupervised and supervised learning. A front-end website was developed to accommodate human interaction. It integrates easy-to-use interfaces for search and identification, detailed and interactive digital representations, and information for educational and motivational purposes. Computer intelligence is used in the back-end to synthesize and leverage human intelligence. To ensure a high quantity of high quality identifications are obtained quickly, the dynamic hierarchical identification algorithm was created to cluster specimens, propagate knowledge, and prioritize input. In this fashion, we provide not only a clear and strong approach to the specific problem of microfossil identification but also a significant case study for image understanding in general.

Acknowledgements

First and foremost I would like to acknowledge my parents who supported me through everything and gave me the drive to always do my best. Without them, I would not have been able devote as much time and focus into my studies. To my brother, thank you for always looking out for me in school, and even more so on campus. A big thanks to my uncle, for being my role model; if not for him I would not have chosen engineering. To my grandmas, thank you for always making me feel loved and spoiled. To my grandpas, I miss you both very much. To all my close friends, thank you for being my emotional support during my years in University.

On campus, Dr. Dileepan Joseph has been a great supervisor and instructor. He has taught me a lot about research, education, technical writing, organization, and enthusiasm about your job. I'm grateful for all the time he has committed to helping his graduate students. He has made my graduate school an enjoyable and productive experience, allowing me to gain a lot of confidence in myself and my work. Dr. Kamal Ranaweera deserves many thanks for all his help getting me started with my server and website; being able to go to him was a great comfort when I first began.

Everyone sharing my lab also played a part in making my experience in graduate school enjoyable. Adam Harrison, thank you for being open and friendly, the first student to make me feel welcome; without all your suggestions and help I would not be where I am now. Orit Skorka, thank you for listening to my constant chatter; your plants helped reduce the sterile, lab vibe. Ali Mahmoodi, thank you for your insights and being a good sport; as you were the butt of many jokes, you helped keep the atmosphere fun and light. And thank you Jing Li, a welcome addition to the lab, the only other student without a coffee addiction.

Contents

List of Tables

List of Figures

List of Acronyms

Chapter 1

Introduction

Artificial Intelligence (AI) covers a wide range of areas, but the most difficult areas in which to conduct research are called AI-complete. An example of an AI-complete problem is image understanding, also called computer vision [1, 2]. It involves programming computers to interpret digital representations by recognizing the objects and settings within them. This interpretation is difficult for many computers because humans take various things into account, such as prior knowledge, context, settings, illumination, and depth [3–5]. In order to advance image understanding, studies focus on a particular topic to define the challenge more clearly; we focus on image identification. The difficulty with image identification comes from the translation of visual features interpreted by humans to features computers can manipulate. Image identification research can be simplified by facilitating segmentation before identification, controlling the images that are used, and selecting an application.

In image segmentation, complexities can arise from the type of images involved. The background and foreground of a photo can have many different features and objects like people walking or buildings. Techniques have been developed to account for the variation seen in backgrounds and foregrounds, such as identification using tree-structured regional features [6], and having context-aware object detection [7].

Simplification of image identification sometimes centers around controlled image capture where the subjects are easily isolated. For example, in fingerprint analysis the method by which fingerprints are obtained is usually controlled. For secure access, a fingerprint scanner is used, and so fingerprints are easily segmented. This leaves the quality of the image to be the limiting factor in the success or failure of identification. Ongoing research is focused on methods to ensure good quality images and to extract object features for processing and identification [8–10].

For AI-complete problems, "solving the problem of the area is equivalent to solving the entire AI problem" [2]. By examining the results for an application of sufficient

importance, the pitfalls and benefits of an algorithm can be better investigated and demonstrated. To this end, marine micropaleontology is an important context that provides a challenging application, namely microfossil identification, with which to investigate the image identification part of image understanding.

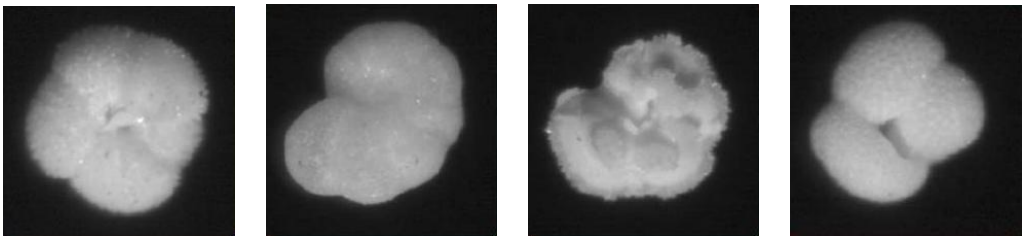
With microfossil identification, image segmentation is simplified using controlled specimen images. Because foregrounds and backgrounds are easily isolated in the digital representations, focus may be placed on foreground identification. The importance of microfossil identification is described in Section 1.1. Human-based computation is a relatively new approach to dealing with AI-complete problems and is described in Section 1.2. A brief description of our approach and the further organization of this thesis is given in Section 1.3.

1.1 Microfossil Identification

Microfossil identification is an important research application because "vistas of new opportunities for pure and applied work in biological and related fields" [11] could be opened. While people are familiar with paleontology, the average person rarely hears about microfossil research and even fewer know the types of microfossils being studied or the significance of this research. For this work, specimens were obtained from ocean drilling programs with a focus on the foraminifera order, which is considered important for biostratigraphy and prehistoric environmental study, and popular approaches to identify Linnaean taxonomy and microfossils in particular were inspected.

Microfossils are readily obtained from three international drilling programs. These programs are the Deep Sea Drilling Program (DSDP) of 1986–1983, the Ocean Drilling Program (ODP) of 1985–2003, and the Integrated Ocean Drilling Program (IODP), which began in 2003 [12,13]. In all these programs, cores were obtained from the ocean floor. Once the cores were extracted and sectioned, they were stored in repositories. Core samples, containing vast quantities of specimens, may be obtained from the repositories for research.

We focus on the calcareous microfossils, ones that mostly or partially contain calcium carbonate, that are widely used in research and industry, in particular, the foraminifera order. Foraminiferida are single cell, shell forming organisms, on the order in size of 1 mm. They are found all over the Earth in oceans and seas [14,15] and date as far back as the Cambrian period [15–18]. Figure 1.1 depicts four foraminiferal "tests" found in our dataset. After death, foraminiferida shells sink to the bottom of water basins where many are fossilized. These fossilized shells are called foraminiferida tests and will be referred to simply as forams. Forams comprise at least 55% and



(a) (b) (c) (d)

Figure 1.1: Foram examples showing four different genera: (a) *Acarinina*, (b) *Anomalinoides*, (c) *Morozovella*, and (d) *Subbotina*.

90% of biomass in the Arctic and deep sea, respectively, and are used to aid with biostratigraphy, paleoclimatology, and paleoceanography [15].

Biostratigraphy is the study of fossils to date rock layers and thereby model the geology of the rock. Foraminifera are appropriate for this type of work because they evolve rapidly [14]. This implies that a particular species is only found in a limited time range. Oil companies find biostratigraphy information helpful in locating hydrocarbon deposits by comparing stratigraphy in different locations [19]. Areas under similar environmental conditions will have similar stratigraphy and foram deposits. Oil companies also use this information to assist with steering drills when drilling horizontally for known hydrocarbon deposits [20]. By constantly checking the stratigraphy in the rock surrounding the drill, oil companies can ensure drilling depth is correct to access known hydrocarbon deposits.

Foram information is also beneficial in the study of prehistoric environmental conditions. In addition to the two spatial dimensions of core location and temporal dimension of age, different foraminifera, benthic or planktic, inhabit different levels of the ocean, yielding a third spatial dimension indicating ocean level when the microorganisms were alive [14]. Forams are useful for geochemical analysis because shell composition is influenced by environmental conditions. When groups of forams are analyzed chemically, the amount of oxygen isotopes, carbon isotopes, and boron present in the forams can be determined [14, 21]. Along with these elements, the analysis of foram magnesium/calcium and cadmium/calcium ratios can help in the reconstruction of prehistoric oceanography and climatology [14]. By studying elements and isotopes inside forams, experts are able to determine prehistoric environmental conditions, such as atmospheric carbon dioxide, carbon cycling, and ocean temperature [14, 21, 22], from local to global scales.

A key step before specimens can be fully used is taxonomic identification. Each specimen must be manipulated and observed under a microscope to determine Lin-

naean taxonomy: order, genus, and species. Manually manipulating each specimen for identification is a time consuming process, limiting the amount of research that may be conducted. Having specialists handle specimens also poses the risk of specimens being lost or damaged. Without another simple and reliable method for identification, manual identification by specialists remains the most popular and reliable method to obtain results. A computerized approach could reduce the amount of physical particle manipulation, allowing many fossils to be identified with less effort and ensuring the microfossils remain available for further use.

Computers are well suited to aid with microfossil identification because they are able to analyze large sets of data quickly. By enabling the identification of many specimens, current studies could advance at a faster rate and other applications may be discovered. Research into microfossil identification varies significantly. Some research centers around image organization like image content-based retrieval [23] and symmetry-based indexing of images [24]. Automatic identification research on microfossils includes many topics, such as: shape identification [25]; automated taxon identification of dinoflagellates [26, 27], diatoms [28], and coccoliths [29]; semi-automatic pollen identification [30, 31]; and other identification of biological particles in microscopic images [32]. The many research papers concentrating on automatic microfossil identification shows the importance and interest in this area.

For our research, we obtain microfossils from established repositories, concentrate on foram identification due to its importance in biostratigraphy and prehistoric environment research, examine popular identification methods, and review the state of the art in microfossil identification. Research specifically for calcareous microfossil identification is divided into two categories: rule-based and ANN-based approaches. Section 1.1.1 describes the state of the art in rule-based approaches and Section 1.1.2 describes the state of the art in Artificial Neural Network (ANN) approaches.

1.1.1 Rule-Based Approaches

Attempts have been made to increase identification accuracy using rule-based approaches. These rule-based approaches focus on assisting experts, specialists, and students to identify microfossils and still require manual particle manipulation. Two of the most prominent systems in this area are Fossil and VIDES.

The Fossil [33] program is an early approach to identification. Fossil attempts to make identification easier using polyclaves, which is a modified version of a decision tree. Users attempt to specify as many attributes as possible, receiving a list of possible taxa when finished. In the ideal case, only one taxon would remain. The

user is able to follow the polyclave search, or specify different identifiable attributes, skipping attributes that cannot be identified. This system searches for the taxon based solely on the attribute values specified so it does not filter results by following the taxonomic hierarchy. This is intended to allow for more flexibility in the system.

Fossil is used by collectors, experts, and students. Collectors are the main users of the system and try to identify specimens by specifying attributes as textual, keyword, or numeric descriptors. Experts provide an attribute template for each specimen taxon and verify each taxon description. Students are also allowed access to the system to test their knowledge by identifying specimen images. In terms of identifying features of a biological specimen, identifying attributes using textual descriptions can be very difficult. This kind of system is geared towards someone knowledgeable in the terms used for accurate identification of specimens. Fossil is more focused towards training students and helping a knowledgeable collector who may not be an expert in specific species traits. Little workload is reduced because of the need to manually specify attributes, meaning the collector would still be required to view and manipulate each specimen under a microscope.

The Visual Identification Expert System (VIDES) [34] assists specialists by reducing training time. This is a visually oriented system, containing images to help with identifications. Care was taken to make this system user friendly, with descriptions, definitions, references, and specific attributes easily accessible. When creating the system, it was noted that the expert being consulted would identify forams by visually identifying features before giving textual descriptions to each feature. With this in mind, images of all attributes were included to assist users with identifying attributes. At higher levels, images are the main information stored to assist with identification because the features are easily distinguished visually.

For all new microfossils entered into VIDES, an expert is required to supply an attribute value table with text descriptions and high quality drawings to be used in identifications [34]. A user is able to specify multiple attribute descriptions to deal with uncertainty, as can be common when identifying biological specimens. The final taxon, or possible taxons, is determined by narrowing the list of available taxons as the attribute list is refined. A drawback is the requirement for a user to be sufficiently familiar with the terminology being used. While this program does have visual images to help with identifications, a user is still expected to have prior knowledge of the attributes and terminology so they are not constantly using the help material. This would not be suitable for users who are not in the given field, limiting the range of users able to perform this task.

Fossil and VIDES have limited benefits for automatic microfossil identification. In

both systems, a knowledgeable user must look at each foram under a microscope and identify attributes, making this an inefficient system in terms of reducing workload for experts and specialists. A desire to fully automate this identification process led to the development of ANN approaches.

1.1.2 ANN-Based Approaches

Developing a fully-automated system is a contrasting approach to developing a rule-based system, which is overly dependent on expert and specialist interaction. There are three popular systems that employ an ANN approach to microfossil identification: CLASSIC, COGNIS, and SYRACO.

One of the first attempts to make a fully-automated identification system used the shell known as CLASSIC [35], a system that shall be referred to simply as CLASSIC. CLASSIC was split into two different processes. The first process collected and analyzed images and the second used a knowledge-based algorithm to identify specimens. Yu *et al.* [35] tried to use foram images obtained from an optical microscope, but it was decided the image resolution was insufficient. In the end, three Scanning Electron Microscope (SEM) photos for each foram, showing three views, were used. In practice, requiring these SEM images to be taken is costly, time consuming, and not practical when identifying large quantities of forams. Obtaining SEM images also have the down side of spoiling the microfossils making them unsuitable for geochemical analysis.

In CLASSIC, major features such as chamber number, chamber shape, ornamentation, and foram boundary shape are all easily identified visually by humans, where different algorithms had to be developed painstakingly for image analysis. In some cases, such as the suture descriptions, textures specific to the forams being tested were used to ensure good results [35]. These taxon specific assumptions make this method harder to generalize. CLASSIC developers also mention they were unable to extract foram umbilical structures. This indicates the difficulty inherent in doing automated feature extraction whereby complex visual characteristics are mapped to simple metrics that are easily processed.

A more recent automated system is called the Computer Guided Nannofossil Identification System (COGNIS) [36]. COGNIS uses a convolutional ANN to analyze SEM images, and the COGNIS Light variation analyzes images obtained using an optical microscope. As SEM images are difficult to obtain and spoil the microfossils, we focus on the COGNIS Light results. Optical microscopes are easy to us and inexpensive, making them a preferred method to obtain images. ANN training was done using

2 000 images and took approximately 30 hours. COGNIS Light was tested using 2 092 images and was only able to achieve good identification results (93%, or 132 images, of *Florisphaera profunda* correctly identified) when the system had a high false positive rate (80%, or 528 images, of other species incorrectly identified as *F. profunda*) [36]. This would imply that final identified results were unreliable with only 20% of the images (132 out of 660 total) in the final *F. profunda* class resulting in correctly identified forams.

Another leading system in automated microfossil identification is called Système de Reconnaissance Automatique de Coccolithes (SYRACO) [29, 37]. SYRACO is an ANN system using four different layers of neurons for identification. When developing this system, face images were used to "test the relevancy of such an approach for pattern recognition of position-normalized objects" [37]. The authors stated they were able to sufficiently generalize the system using only 200 training images. As a final note, the authors wrote "it is still difficult to explain how a network with about 800 000 free parameters and trained on 200 images of faces can correctly identify 91% of unseen faces" [37]. The difficulty in justifying their results is a drawback that would prevent further expansion and, for some researchers, confidence in the approach.

CLASSIC, COGNIS, and SYRACO are all ANN systems developed to perform automated microfossil identification. These previous systems have met with limited success. The problems with the rule-based approaches center around the high amount of work required of knowledgeable users. ANN-based approaches are also limited because they depend on difficult-to-obtain SEM images, encounter high incorrect identification rates, or are difficult to understand and, therefore, justify and build upon. For further advancement in microfossil identification, a new approach centered on human-based computation, is used.

1.2 Human-Based Computation

Human-based computation is a new approach to tackle AI-complete problems and has not been used previously with microfossils. With human-based computation, a computer outsources tasks to people. Humans are desired to aid with image understanding because they are capable of robust and efficient visual analysis. The difficulty in programming feature extraction, the wide range of potential taxons, and the natural variability within a given species add complexity to automated identification. However, computers are exceptional at processing large amounts of data quickly. Human-based computation attempts to leverage the benefits of both human and computer processing. Human-based computation also has strong ties to many

other terms that fall under "crowdsourcing". To better understand crowdsourcing, it is important to understand how the terms are related and defined.

There are many terms related to human-based computation and some definitions are still evolving. These terms include crowdsourcing, citizen science, citizen cyberscience, and distributed thinking. All these terms imply getting input or information from a group of people, usually volunteers. The differences in these definitions are in the scope they cover. A visual representation of the overlap between these different terms is shown in Figure 1.2.

Crowdsourcing is the process of requesting help from a large number of individuals to complete a task, usually for free. Crowdsourcing is an all encompassing term and includes groups of people with or without previous training, paid or unpaid work, scientific or unscientific work, and computer-based or non-computer-based work.

Citizen science is the use of volunteers to assist with scientific research from data collection to data analysis [38]. This area may or may not use computers for research. Citizen science projects also educate the public about the scientific process and the particular field being studied. Citizen science is quickly gaining popularity. One website called Science for Citizens (`scienceforcitizens.net`) is devoted to connecting volunteers with citizen science projects around the world [39]. This site currently has 223 projects included in their database and it is safe to say many more are not included or still under development. Citizen cyberscience is a subset of citizen science where computers are more essential to the research; in many cases, the Internet also plays an integral role in the research.

Distributed thinking is focused on multiple people, in a community of computer users, performing computer tasks to achieve a common goal. Distributed thinking is very similar to citizen cyberscience. However, the work being conducted by distributed thinking does not need to be used for scientific research.

Human-based computation is defined more from the purpose of the project. Projects that employ human-based computation are focused on the program attempting to perform a task or reach a goal and being unable to perform certain tasks. The difficult tasks are outsourced to humans. In human-based computation, the computer does the majority of the work with people performing smaller tasks enabling the computer to continue processing to reach the final goal.

With the many terms covering a variety of different scopes, and with sometimes changing definitions, it may be difficult to categorize a given project by just one of these terms. In many cases, projects involve multiple aspects enabling them to be categorized under multiple definitions. For example, Foldit [40] is considered a citizen cyberscience and distributed computing project. Distributed computing is

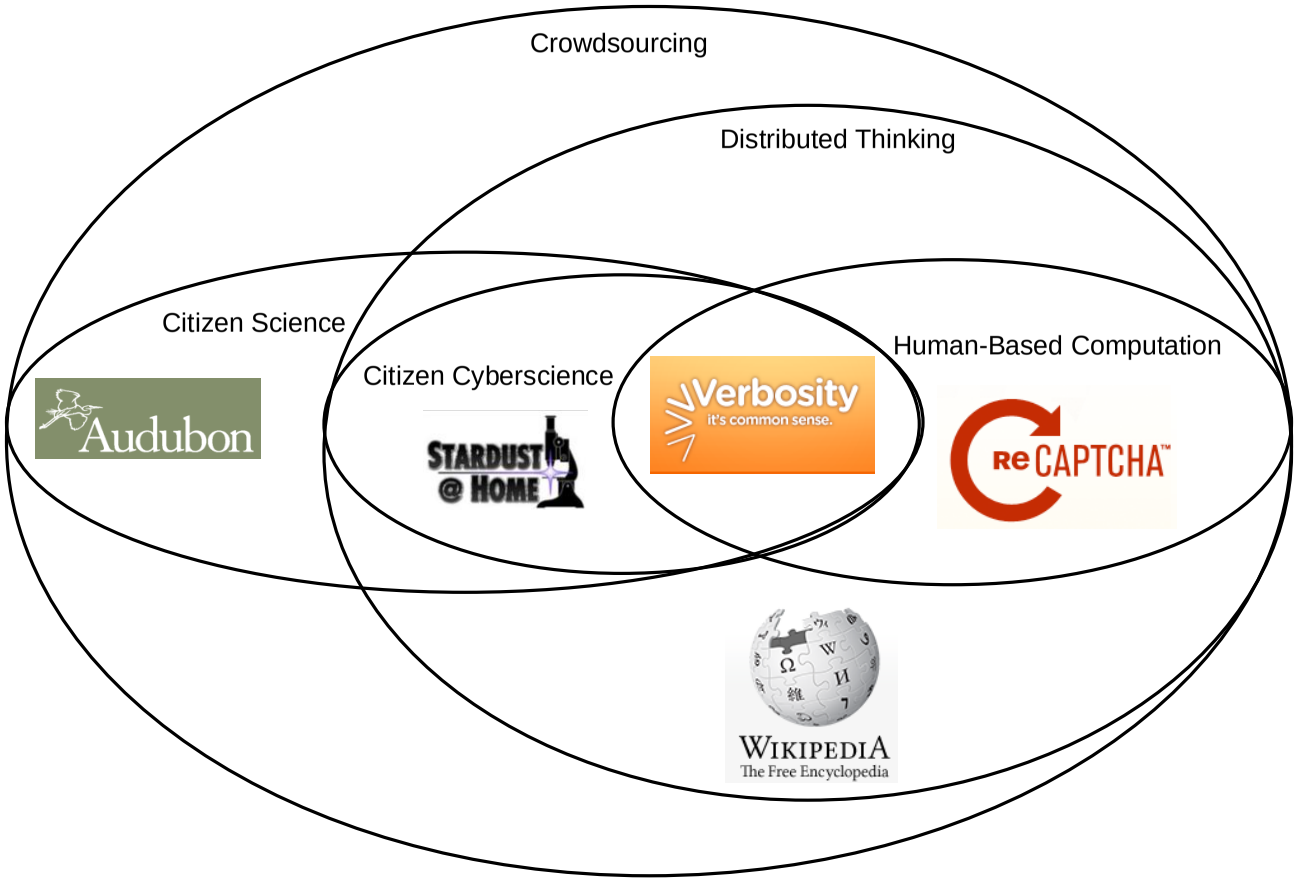


Figure 1.2: Venn diagram depicting the definition overlap between common terms used to describe outsourcing tasks to a large number of volunteers. Notable examples are given.

very similar to distributed thinking—all tasks are programmed but one computer is unable to perform all the operations so tasks are farmed out to other computers volunteered by human users.

Human-based computation is the main approach now being taken in our research and it falls under the crowdsourcing umbrella, which includes many evolving terms, each one associated with a number of different projects. Instead of attempting to make sense of the state of the art by looking at a specific term, crowdsourcing projects are grouped based on their level of human versus computer involvement in the completion of the project. Figure 1.3 shows an axis indicating the level of human versus computer involvement in well-known projects. If the axis is bisected into two sections, the projects can be split into human focused and computer focused groups. The human focused projects are described in Section 1.2.1 and the computer focused projects are described in Section 1.2.2.

1.2.1 Human Focused

Human focused projects require a relatively large amount of human input. The amount of human focus may be determined by the final goal of the project. This project range begins with the Christmas Bird Count [41–43] where humans do most of the work, such as the data collection and analysis, while computers merely store the information. As we move along the axis, computers are more involved in the isolation and collection of the information. Examples of these projects are Herbaria@home, ESP Game, Stardust@home, and Galaxy Zoo.

The popular annual Christmas Bird Count is the most human focused project. Starting in July 1900, the American National Audubon Society (ANAS) held the Christmas Bird Count for approximately two weeks going across December and January [41]. Volunteers monitor and catalogue birds observed from various locations in the wild. From December 2007 to January 2008, there was a total of 59 918 volunteers who counted a total of 57 704 250 birds [42]. In addition to this, ANAS has started the Backyard Bird Count, which has been running since 1998 and is held for about four days in February [43]. In 2011, they received 92 206 checklists with 11 471 322 birds counted comprising 596 species [44].

Herbaria@home attempts to identify the museum and university herbarium collections in the United Kingdom, of which there are about 20 million herbarium specimens [45]. At the website, volunteers are shown a high quality image of each herbaria sheet. They try to decipher the documentation in the sheet by analyzing the writing or symbols shown on the image. The project fully documented 40 000 specimens and

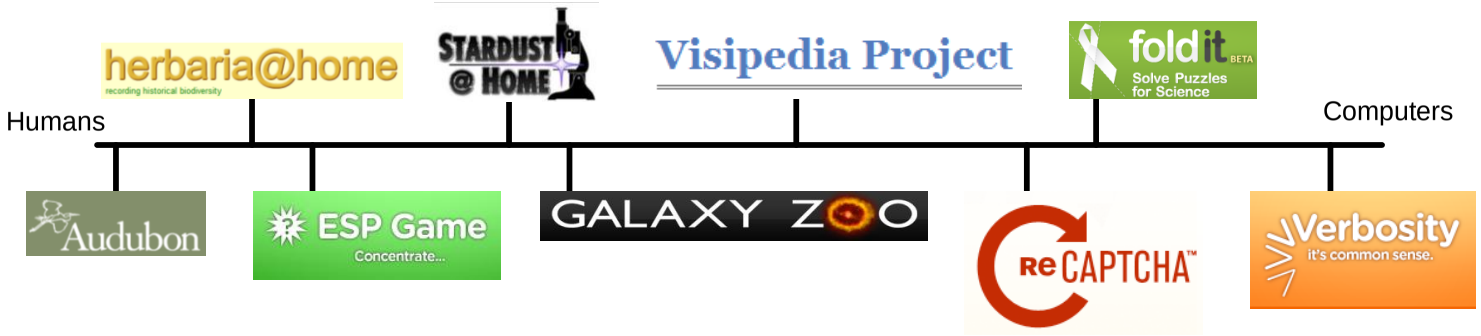


Figure 1.3: Axis indicating the level of commitment from humans versus computers in well-known crowdsourcing projects. The placement of each crowdsourcing project on this axis shows the current state of the project relative to the others.

80 000 sheets in the first three years and, as of December 2010, the website indicates 74 880 specimens have been documented [45]. Compared to the Christmas Bird Count, this project has slightly more computer processing through the digitization of the herbaria sheets for users to view. In this case, data analysis occurs by looking on a computer instead of in nature.

The next project on the human versus computer axis is the ESP Game [46, 47], named after extrasensory perception. This is an online game designed to obtain labels for web images because "there are millions of [pictures and] no guidelines about providing appropriate textual descriptions" [47]. Having labels for the many images found on the web makes images easier to search. Obtaining image labels could also be further used "as training sets for machine learning algorithms" [47]. At the current stage, the specific design for this system does not actively feed results into other machine learning algorithms and instead focuses on collecting the labels for searches.

The novelty of the ESP Game, when it was first released, is the attempt to convert the data collection into a computer game to entertain users and motivate them to keep playing. The game itself is implemented as a Java applet [48]. Based on several experiments, a pair of players could produce on average 3.89 labels per minute [47]. Between August and December 2003, 13 630 people played and were able to generate 1 271 451 labels for 293 760 images. It was also noted that 33 players played more than 1 000 games with over 10 904 players playing more than once. Results from this game has spawned Games With A Purpose (GWAP), which are systems that collect information through game play.

Stardust@home "is one of the pioneering distributed-thinking projects" [49]. It is funded by grants from NASA's Science Mission Directorate and is run by the Space Sciences Laboratory at the University of California at Berkeley [50]. This project asks volunteers to help locate interstellar dust tracks collected from the Stardust return capsule. For this system, high quality images are used along with JavaScript to create a virtual microscope that a volunteer can control with her mouse to see different levels of tile focus [51]. Users are asked to locate dust tracks in aerogel tiles and specify the location of the track by clicking on the image. Aerogel tiles are low density, silicon-based, solid tiles containing 99.8% air [52].

Volunteers were eager to participate in Stardust@home and would actively engage in discussions on a forum, where they named themselves 'dusters'. In total, 20 064 dusters participated, from August to December 2006, and were noted as contributors in a resulting paper [51, 53]. The greater processing required to take all the high quality images and create the virtual microscope means this project has more computer involvement than the ESP Game.

Galaxy Zoo is a popular project enabling volunteers to assist researchers by labeling galaxies photographed by the Sloan Digital Sky Survey [54, 55]. In the first paper, raw data collected over a five month period had 85 276 users giving 893 212 galaxy identifications. After one year, the number of participants grew to 150 000 volunteers, making 50 million identifications [55]. When analyzing the final data, removing repeated identifications resulted in around 39 distinct identifications per galaxy [54].

Overall, Galaxy Zoo was well received by volunteers and shows how crowdsourcing projects can work successfully to identify images for further analysis if carefully planned. Galaxy Zoo is just after Stardust@home on the human versus computer axis because images are taken, galaxies are isolated, and then are displayed to volunteers automatically through an interactive, visually appealing, and dynamic user interface. Further analysis of the human input is performed by computers, with interpretation of results conducted by humans. This is why Galaxy Zoo is still considered a human focused project, but one that is near the middle of the human versus computer axis, and which borders the computer focused projects.

The Christmas Bird Count, Herbaria@home, ESP Game, Stardust@home, and Galaxy Zoo are examples of human focused projects. The main focus and goals of these projects require relatively more human input when compared to the remaining examples in the computer focused range on the human versus computers axis.

1.2.2 Computer Focused

Compared to human focused projects, computer focused projects involve more processing or automated tasks. Computation in these projects requires more automation and less human involvement over the ultimate scope determined by the project. Computer focused projects include Vision of Visipedia, reCAPTCHA (a project based on CAPTCHAs), Foldit, and Verbosity.

Visipedia is a work in progress to develop a "visual interface for Wikipedia that is able to answer visual queries" [56]. Visipedia developers want to create a system allowing experts to upload and label images. New labels would only be required for any sections in the image not previously labeled, with all the other common sections being automatically recognized and labeled using previous image labels. This workload reduction is more desirable for experts. Users, editors, non-experts, and annotators would be using the system, in addition to the experts who enter information, check information, and use the information.

The scope of the completed Visipedia project is very broad and requires a lot of automation. In order to progress to the final vision for the Visipedia system,

the developers started with the first system by picking ornithology, "a well-defined domain... with a community of highly motivated enthusiasts" [56]. The current system is quite dependent on human interaction. However, the final system will eventually become highly automated as suitable algorithms and image understanding techniques are developed to relate and label similar objects detected within images.

reCAPTCHA is a modification of the Completely-Automated Public Turing Test to tell Computers and Humans Apart (CAPTCHA). A CAPTCHA is a test used online to determine if a human or a computer is making a request [57]. A normal CAPTCHA is a group of characters that has been distorted to an extent that computers cannot recognize the symbols although humans can. They are used to prevent the abuse of services that are offered online such as free email [58]. Ahn *et al.* estimated around 100 million CAPTCHAs are typed every day [57]. reCAPTCHA was developed because "deciphering CAPTCHAs requires people to perform a task that computers cannot" and the authors wanted to put that effort to further use, namely to digitize text [57].

The reCAPTCHA project is computer focused because it first attempts to automatically read text from scanned material. If two Optical Character Recognition (OCR) programs do not agree on the word, it is converted into a reCAPTCHA by pairing it with a known control word and using the result like a CAPTCHA to test if a human or computer is attempting to make a request online. Preliminary reCAPTCHA testing showed word level accuracy to be 99.1%, whereas OCR systems typically have 83.5% accuracy [57]. The benefit of reCAPTCHA is better understood if CAPTCHA success is considered. Ahn *et al.* wrote that humans solved over 1.2 billion CAPTCHAs in one year, which means over 440 million words could be deciphered by users if reCAPTCHA was used instead [57]. If you assume a book holds 100 000 words then this would imply the potential to transcribe 176 000 books per year. Because the final goal of this project is to translate text, and much of this process is already automated, this system is next on the human versus computer axis. In this case, humans are only used to decipher the limited remaining text that was not recognized by the OCR software and so they are used to deal with the exceptions and not the typical word going through the system.

A more advanced project is Foldit, which was developed after the success of Rosetta@home, a distributed computing system created to do protein folding with a screen saver that allowed people to visualize the progress [49]. Feedback received from Rosetta@home volunteers showed that people watching the screen saver would get frustrated with the time the computer took to process the many degrees of freedom as needed to arrive at an optimum energy state when they could see a better

state. Human ability to see the better state is due to our "highly evolved talent for spatial manipulation" [49]. As a result, Foldit was developed.

Foldit is "a multiplayer online game that engages non-scientists in solving hard prediction problems" [40]. The goal of this game is to create "accurate protein structure models" [40]. Due to the added computer processing involved in the distributed computing stage, this is a highly computer focused system. The main reason why Foldit is not the most computer focused system out of all the chosen examples is because humans are asked to perform tasks that are equivalent in complexity to tasks the computer is required to perform. In the overall system, the computer is able to complete most of the tasks. However, humans still have a big impact on results as they are able to perform the same kind of analysis faster and, in some cases, better than the computer. For this reason, Foldit is ranked behind Verbosity as the most computer focused project.

Similar to the ESP Game, Verbosity is also a GWAP system. Whereas the ESP Game was used to label images only, Verbosity is designed to have results taken and used to develop an intelligent system for natural language understanding [59]. Verbosity is used to collect common-sense facts by involving two players. One player is shown a word and must help the second player guess the shown word. The first player is given the option of completing one or more of six hints [48, 60]. The second player attempts to guess the word using the hints. In the end, it was determined that results from GWAP systems are "bound to be somewhat noisier" than data from the established ConceptNet database, generated using data collected from the Open Mind Common Sense (OMCS) project [60,61], due to the nature of volunteers participating for personal enjoyment in the game and not to assist scientists by providing data for computer science applications.

Verbosity is considered the most computer focused project because the project, while complete, is only considered as one stage towards the main goal, which is to develop a general-purpose intelligent system by incorporating these common-sense facts. The collection of common sense facts from humans is minor in comparison to the major goal of an intelligent system that understands language by taking in, interpreting, and responding to input automatically. Consequently, this project is the most computer focused out of all the examples given.

Vision of Visipedia, reCAPTCHA, Foldit, and Verbosity are four examples of computer focused projects. The wide range of successful human and computer focused projects show volunteers are willing and eager to participate in crowdsourcing projects. Despite the wide range of tasks, difficulty, and motivating factors, data collected by these works demonstrate successful if somewhat noisy results, which depend

on the information collection approach. For many of these projects, the tasks performed by the volunteers are broken down to ensure simplicity and a quick learning curve while keeping volunteers motivated. These are important considerations when creating a similar system to tackle microfossil identification.

1.3 Scope of Thesis

Through the incremental development of a fully-automated microfossil identification system, not only do we advance microfossil identification, but also we contribute to the advancing of artificial intelligence. Due to the complex nature of microfossil identification, we have taken the approach of breaking the problem down into smaller, incremental steps. As a major step towards our final goal, this thesis proposes a human-based computation system using citizen cyberscience to reduce the need for specialists and experts. Although our crowdsourcing approach involves multiple facets, we emphasize its goal of full automation by calling it human-based computation. This thesis provides a detailed description of our design methods and the components developed for this application.

The rest of this thesis is organized into four chapters. Due to the complexity of the system, we use an evolutionary prototyping approach to design. A description of previous prototypes and the design cycle for the current prototype is given in Chapter 2. As the human interface is a major component of this system, a website developed for citizen cyberscience is described in Chapter 3. This chapter details design considerations and layout, while also presenting the text currently found on the website. The website itself is only the front-end of the human-based computation system. Chapter 4 elaborates on the back-end by detailing the computation algorithms we have designed, programmed, and tested. The last chapter, Chapter 5, concludes with a summary of thesis contributions and proposes future work for the next stages of development.

Chapter 2

Evolutionary Prototyping

To advance research in microfossil identification, we want to create a reliable, accurate, and affordable method to identify microfossil specimens automatically. However, reliable automatic identification for a wide range of taxons is very difficult due to natural variability within species. In order to develop a completely functional system, it was decided to approach this problem using a design life cycle model. There are several types of software development life cycle models that can be used. The waterfall model, spiral model, v-model, throwaway prototyping model, and evolutionary prototyping model [62,63] are all well known life cycle models in software engineering. These models describe the various project stages such as elicitation of specifications, requirements engineering, initial design, implementation, testing, and maintenance. When examining the different life cycle models, evolutionary prototyping was chosen because it is ideal for academic time constraints, crowdsourcing, and exploratory research.

In evolutionary prototyping, once the general specification of desired functionality is outlined, a prototype is developed and modified iteratively depending on the results obtained from previous prototypes. This design model reuses, modifies, and expands on previous prototypes unlike other models that develop the full system in one stage or throw away earlier prototypes and start from the beginning for each new prototype. The evolutionary prototyping model, shown in Figure 2.1, was used because of the novelty of our microfossil identification approach. The new crowdsourcing approach makes it necessary to view results, examine performance, and verify desired functionality. At each iteration, a new functional prototype is required to generate results for analysis. It is expected that new issues, considerations, and ideas will also appear through the course of the system design. By reusing and modifying previous prototypes, we are able to save time as we do not have to rebuild the system every time.

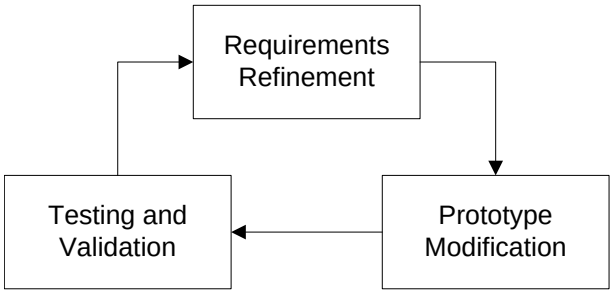


Figure 2.1: The design life cycle for the evolutionary prototyping model. A single prototype cycle consists of a requirements refinement phase followed by implementing the necessary prototype modifications before performance is tested and expected behaviour is validated.

Other systems use the evolutionary prototyping design life cycle even if it is not explicitly stated. Visipedia is a good example of an evolutionary prototyping system because the developers use an incremental approach. The ideal finished system is extremely complex, requiring more developed automation compared to the current state of image understanding. Visipedia requires not only objects (like birds) but also parts of objects (like beaks) to be segmented, object relationships to be recognized (like beaks on all birds), image segments to be hyper-linked autonomously, and all of this to be done on any uploaded image. The approach being taken for this project is to add "useful automated agents, one by one" [56], an evolutionary prototyping stance.

With crowdsourcing systems in particular, predicting how the general public will react is difficult. This unpredictability is compounded because the desired participants include a wide range of ages and experiences. Galaxy Zoo is an example of an evolving crowdsourcing system where participation was unpredictable. Through the course of their study, participation exceeded expectations. Due to the success of the initial pilot project, Galaxy Zoo continues to evolve and has now developed into the Zooniverse system, which incorporates Hubble Telescope images and various astronomical elements such as galaxies, merging galaxies, supernovas, planets, solar storms, and lunar surfaces [64]. The content in Galaxy Zoo itself has also evolved. The original project had simpler identifications whereas the current system asks users for more specific information. For example, the pilot study asked if galaxies looked elliptical; now, they ask if elliptical galaxies look circular shaped, cigar shaped, or somewhere inbetween [55].

Similarly, Foldit also evolved after trial runs and user input. The investigators "fine-tuned the game through continuous iterative refinement based on observations of player activity and feedback" [40]. Foldit has gone even further after trial runs

yielded exceptional results and allowed researchers to learn new strategies developed by successful players such as recognizing that a period of energy increase could ultimately result in lower final energy. The system itself was gradually refined based on player activity, observations, and feedback. Foldit players have proven their ability to manipulate proteins and developers have incorporated additional options for volunteers to design new proteins that could assist other researchers, such as proteins that could bind to pathogens such as HIV [49]. Because of the unpredictability of humans, evolutionary prototyping may be a best practice for crowdsourcing projects to adapt to human strengths and to learn from human strategies.

Developers who create systems for research also use prototypes. In exploratory research, systems are modified and refined partly through trial and error. In many cases, prototypes are not mentioned or included in publications as they were stepping stones used to obtain the final system reported in the publications. Sometimes, prototypes are described in publications as a system evolves, as with, for example, the Artificial Neural Network (ANN) systems Système de Reconnaissance Automatique de Coccolithes (SYRACO) 2 [37] and SYRACO [29]. An interesting note is that SYRACO 2 was published first. As the 2 stands for version 2 [37], it also implies a version 1, an earlier prototype, that is not explained in the publications. In a later publication, the SYRACO system is mentioned without a version, but the authors state that they "present a modified version of SYRACO" [29], implying that SYRACO is another prototype of a system developed by improving upon SYRACO 2.

Compared to all the other software design models, evolutionary prototyping is optimal for academic time constraints, crowdsourcing, and exploratory research. The reuse of previous prototypes reduces the amount of time spent developing improved systems. Public reaction is difficult to anticipate with crowdsourcing, especially for a new project, in a new area, and when introducing a variety of different elements. Due to the nature of research and having to test and analyze results, evolutionary prototyping approaches are often used, if not explicitly stated. This leads a strong preference to build a system using the evolutionary prototyping design life cycle.

This chapter describes the Microfossil Quest system and how it was developed through evolutionary prototyping. Section 2.1 describes the previous prototypes that led to the current system. The design cycle for the current prototype is then described in more detail. Section 2.2 presents a description of the requirements identified for the current prototype. After the requirements were understood, the previous prototype was modified and new components were implemented as explained in Section 2.3. The last stage of development for the current prototype is the testing and validation phase, covered in Section 2.4.

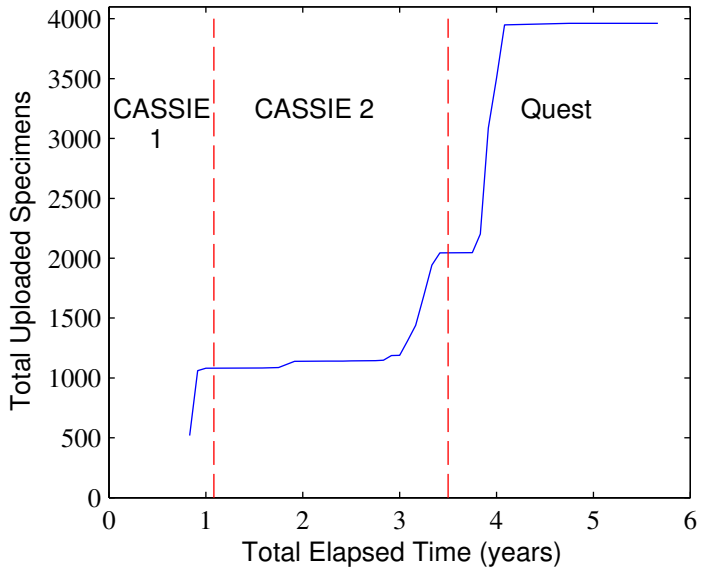


Figure 2.2: Evolutionary frame of the first three prototypes of the Microfossil Quest according to when specimens were uploaded to the database. Time zero represents January 1, 2006, when the research program started.

2.1 Prototype Evolution

In our work to advance microfossil identification, we designed a system now called the Microfossil Quest. At this point in time, the Microfossil Quest has gone through two completed prototypes and a recently completed third prototype. Looking at our database, an outline of the time taken to develop the three prototypes can be observed before details are given for individual prototypes.

The prototype design stages can be divided according to the dates specimens were loaded into the database. In most cases, the loading of new specimens indicates the end of the implementation or prototype modification phase and the beginning of the testing and validation phase. A plot of the total number of specimen ID values in the database over time is shown in Figure 2.2.

In general, a single specimen is given one specimen ID in the database, so Figure 2.2 gives an indication of the number of specimens entered into the database. From the beginning of the graph to the first vertical line, at year $1\frac{1}{12}$, indicates the development period of the first prototype, called the CASSIE 1 prototype. The development period for the second prototype, named CASSIE 2, is between the vertical line at year $1\frac{1}{12}$ and the vertical line at year $3\frac{1}{2}$. In both cases, a major increase in specimen IDs delimits the prototype time frame. From year $3\frac{3}{4}$ to year $4\frac{1}{12}$, there is another large increase in total specimen IDs. However, in this particular case, a separate study was being conducted on potential improvements to the system. Results

from this study play a minor role in the third prototype, called the Microfossil Quest prototype. As these increases were done during the design of the third prototype they are also included inside the time period for the Microfossil Quest prototype, which is from year $3\frac{1}{2}$ to about year $5\frac{2}{3}$, the present.

To obtain a firmer grasp of how the system has evolved over time, we go into further details on individual prototypes. Going into more detail to understand our ultimate goal, the final system is first outlined in Section 2.1.1. Our reasons and considerations for each prototype design cycle is then outlined. In the first prototype, described in Section 2.1.2, research began into creating a computer-aided system called CASSIE. The second prototype, outlined in Section 2.1.3, incorporated a major improvement to CASSIE and is called CASSIE 2. The reasoning behind the third prototype, the Microfossil Quest, is described at a high level in Section 2.1.4. Detail on the Microfossil Quest prototype requirements, implementation, and validation is given in Sections 2.2 to 2.4.

2.1.1 End Prototype (Commercial System)

The long-term goal is to develop automated microscopy and intelligent system technologies to enable detailed biostratigraphy. With the incremental development of the system, we target the ultimate goal of a commercial system. Our end prototype, also called our commercial system, will be a sufficiently-automated microfossil identification system. The commercial system would contain hardware products and a software service that would be available to industry and researchers. The hardware would incorporate the necessary equipment for specimen acquisition. The software service would quickly identify a high quantity of specimens while maintaining high quality identifications.

2.1.2 First Prototype (CASSIE 1)

To incrementally create the end prototype, it was decided the first prototype should focus on the hardest issues arising from existing rule-based and ANN-based approaches to microfossil identification. To decide and develop the first prototype, we reviewed the state-of-the-art in microfossil identification and determined the approach our system would use, refined the requirements, implemented the system, and validated system results.

A review of the state-of-the-art in microfossil identification included an examination of rule-based and ANN-based approaches. Rule-based systems normally require a knowledgeable user familiar with terms regularly used to describe different tax-

ons. The major drawback to rule-based systems is the need for a specialist to view and manipulate each specimen under a microscope, which remains time consuming. ANN-based systems, such as the Computer Guided Nannofossil Identification System (COGNIS) Light and SYRACO, were able to obtain 93% and 91% correct identification rates respectively [36, 37]. However, they required difficult-to-obtain and specimen-spoiling Scanning Electron Microscope (SEM) images, were unreliable as indicated by high incorrect identification rates, and/or their ability to generalize was questionable, which made them difficult to expand and justify.

Instead of a rule-based or ANN-based approach, we decided to create a computer-aided approach. This new method sought to reduce the workload of experts while maintaining the accuracy of taxon identification by experts who are more reliable than computers. This first prototype was created to assist experts by reducing their workload as opposed to fully identifying specimens automatically.

In the requirements refinement phase, we determined the system would digitize, cluster, and identify specimens. Once digitized, specimens were entered into a database where a computer program clustered the specimens. After clustering, a template from each cluster was automatically selected for identification. To maintain accuracy and avoid the difficulties of isolating specimen features, a specialist or expert looks at the cluster template and identifies the specimen using the particle itself or its digital representation. The former must be done with a microscope; the latter may be done online using a website called the Microfossil Wiki. The computer-aided system assists specialists and experts by reducing their workload through clustering and template selection, as any specimen inside a cluster is given the same identification as the template.

The first prototype was implemented by Ranaweera; there was no pre-existing system to modify. This first system is called the Computer-Aided System for Specimen Identification and Examination (CASSIE) 1. The design of the CASSIE 1 prototype is shown in Figure 2.3. There are three main components in the system: specimen acquisition, computation algorithms, and human interaction. In the specimen acquisition component, physical specimens are manually placed under a microscope where C++ software is used to capture images, and to upload the images and associated specimen data to a database.

In the computation algorithms component, the preprocessing and clustering of specimens is completed. Preprocessing of the digital representation was completed to ensure all specimens are oriented the same way and must be done on all digital specimens before use. MATLAB was used to normalize specimen images through the use of an invariant transform based on principal component analysis, with am-

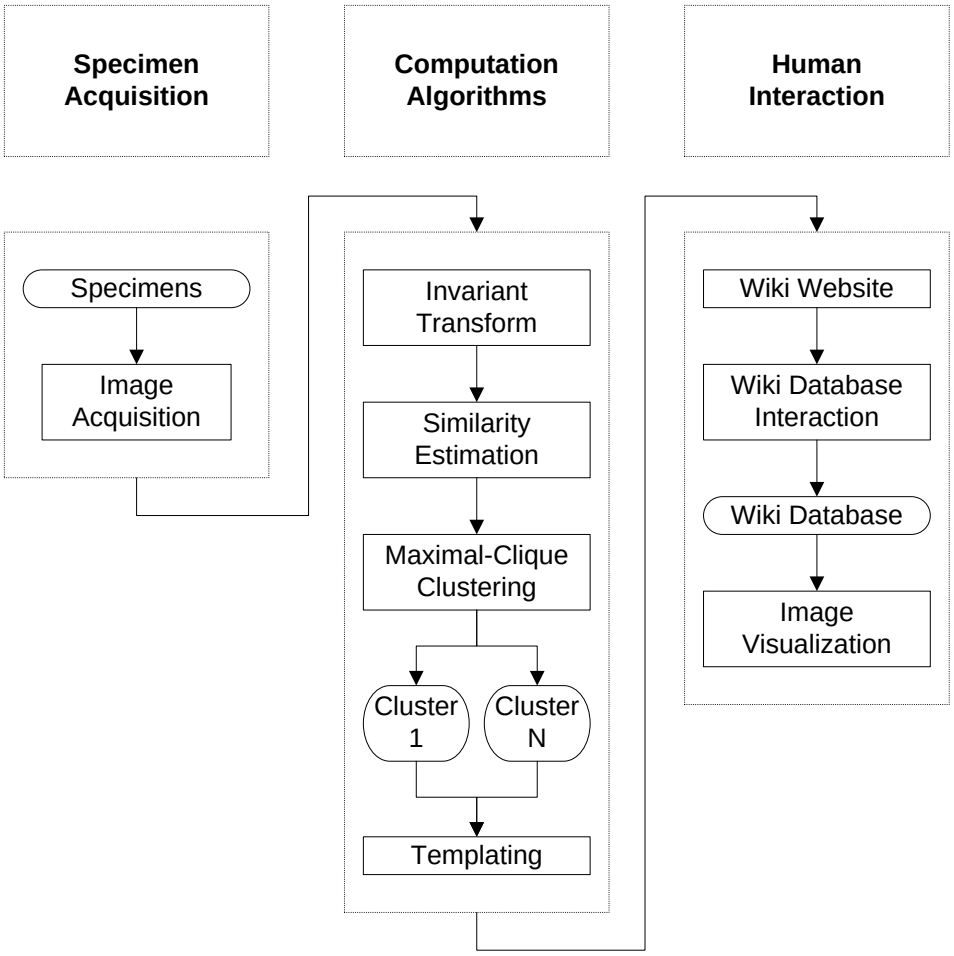


Figure 2.3: Breakdown of system components and data flow for the CASSIE 1 prototype. Specimen acquisition captures specimens manually placed under the optical microscope. Computation algorithms contains the preprocessing performed on the digital representation along with specimen clustering. Human interaction was developed as an external feature, the Microfossil Wiki website, to obtain expert identifications.

biguities resolved using the largest third-central-moment of the specimen [65]. After preprocessing, the computation algorithms component performs similarity estimation, clustering, and templating of digital specimens, all of which was coded in MATLAB. The similarity estimation used correlation coefficient ratings while maximal clique clustering with a threshold was used for clustering [65]. From the clusters, a template was selected by taking the specimen with the greatest similarity to all other images in the cluster. After the template was identified, all other specimens in the cluster were given the same identification. This identification method was called Practical-Template Clustering (PTC).

For the initial system, we wanted to allow our volunteer specialist to provide direct

identifications from anywhere in the world. For this reason, the human interaction component uses a website called the Microfossil Wiki [66], or Wiki for short. In the CASSIE 1 prototype, the Wiki, website database, and interactions were developed with the intent that only a few users would be accessing the Wiki to provide identifications. The site and its data would remain open to public view, but only a few experts would provide identifications. Specimen visualization was very basic with the use of images for this prototype. The only visualization options were a single view of the specimen, or two views. The single view shows the default image taken after randomly dropping specimens on a slide. For the two views, the default image is used and another image is taken after the specimen is manually overturned to show the opposite side of the specimen.

The database itself was created using MySQL to store the information. The website interacted with the database through the use of PHP and MySQL commands. MATLAB software interacted with the database through the extraction of database results that were then imported into MATLAB.

After the implementation of the system, testing and validation was conducted. The testing methodology and specimens used are described by Ranaweera *et al.* [66]. For the main set of results, 238 genus identifications and 169 species identifications were obtained. This dataset yielded a 81% Correct Genus Rate (CGR), a 47% Correct Species Rate (CSR), a 4% Incorrect Genus Rate (IGR), and a 4% Incorrect Species Rate (ISR) when comparing Image-Based Identifications (IBIs) and Particle-Based Identifications (PBIs) [66].

When comparing PBI, IBI-1 (single view), and IBI-2 (double view), it was shown that "while the availability of alternative-view images led to improvements [in identifications], they were small" [66]. This led to the conclusion that the effort needed to manually manipulate specimens for multiple views was not reflected in performance improvement. In terms of image quality, assumed to include sharpness and illumination, a high impact on identification results was detected with a CGR of 96% and a CSR of 63% for good image quality ratings [66].

Results of the PTC method using PBIs yielded 90% correctly identified specimens with a 35% reduction in relative effort and an incorrect rate of 5% [65]. This reduction in workload was due to specialists and experts only having to provide identifications for templates and not every single specimen. Similarly, using IBIs, a comparable performance to 100% effort is seen up to a 35% reduction in effort. It was noted that illumination direction had a high impact on image variability [65] and, therefore, clustering results. Ranaweera *et al.* mentions "a low similarity score is possible, even for a pair of images of the same specimen" [65], if the illumination directions

are different. Vital information concerning specimen shape can also be seen when altering the illumination direction, and it "presents a clearer picture of [specimen] morphology" [65].

Once the system approach was decided, the requirements refinement, implementation, and testing phases of the CASSIE 1 prototype was conducted. It was determined that digital images taken with an optical microscope provided sufficient information for a knowledgeable user to provide accurate identifications. The illumination direction present in the images had a high impact when clustering microfossils. For this prototype, there was also no automation in the process to obtain the digital representations so all specimens had to be located and manipulated manually, which was seen as very tedious and time consuming. The desire to improve automation and digital representations led to the second prototype.

2.1.3 Second Prototype (CASSIE 2)

For the second prototype, focus was placed on improving the CASSIE 1 prototype. The main issues discovered were problems due to varying illumination having a large impact on results. The CASSIE 1 prototype also encountered difficulty because of the manual manipulation required to obtain digital specimen representations. The manual manipulation transferred the tedious work of manipulating specimens from a user providing identifications to the user entering digital specimens into the system. To address and obtain many illumination angles for a single specimen, an autonomous image digitization component became important. This led to requirements refinement and prototype modification, which created the second prototype.

The major requirement in the second prototype was to obtain better digital representations by taking specimen illumination into account. For this reason, the second prototype is called CASSIE 2. To examine illumination, the system requirements for the second prototype were first to reduce the tedium of having to manually locate and capture each specimen. With the ability to automatically take multiple pictures of each specimen, multiple images under different illumination conditions could be taken and examined. By allowing users to view the specimen under varied illumination, we hoped to improve identification results with more accurate and informative digital representations.

The CASSIE 1 prototype was modified in order to incorporate the new system requirements for CASSIE 2. The modifications were made by Harrison, with the final system flow shown in Figure 2.4. Previous components of the CASSIE 1 prototype were modified with a new component added for specimen dissemination.

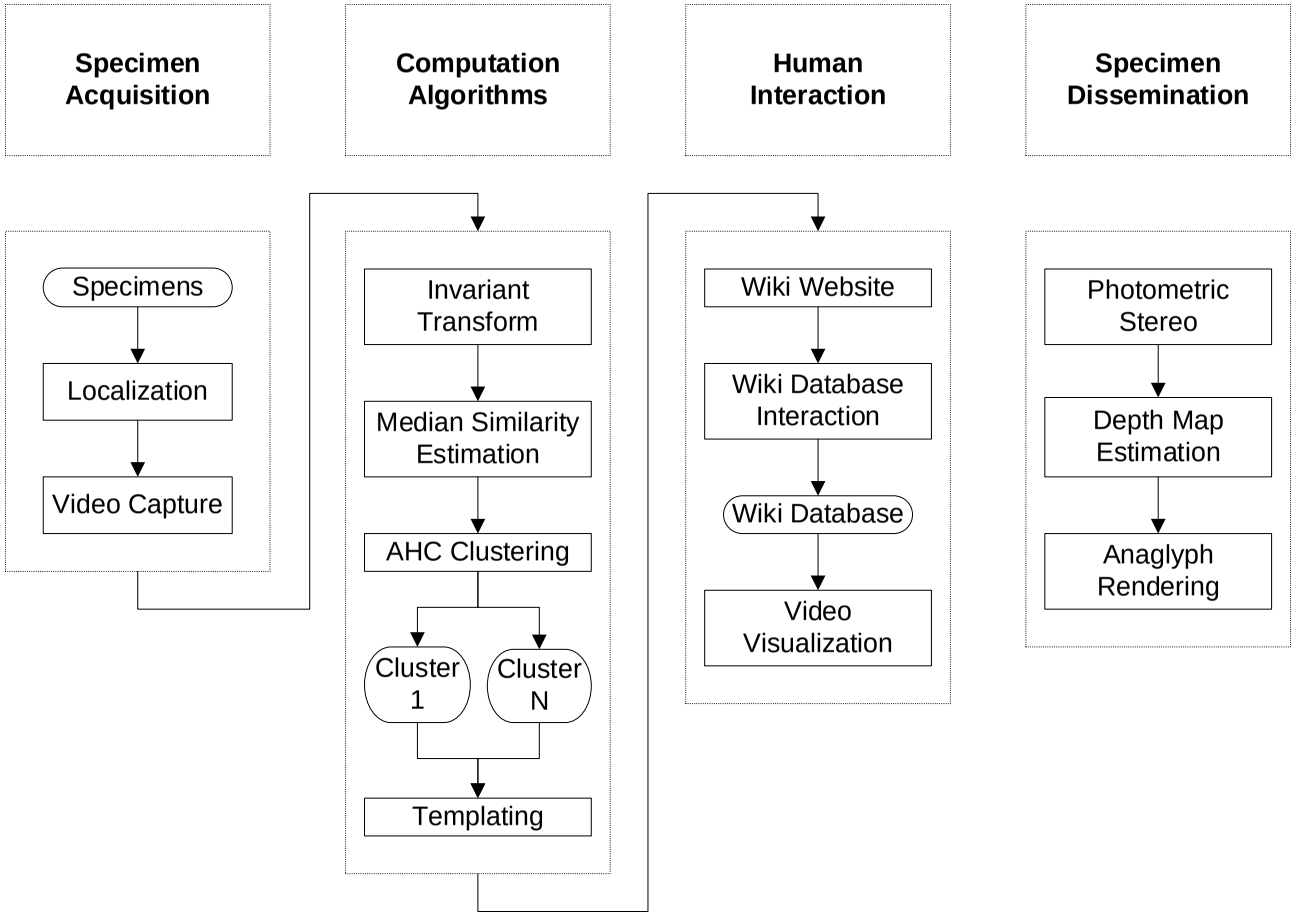


Figure 2.4: CASSIE 2 components and data flow. The main system modifications were implemented in specimen acquisition and computation algorithms, with minor revisions in direct identifications. Preliminary research was also conducted into anaglyph (3D) representations and is included in a specimen dissemination component.

Specimen acquisition was modified by incorporating automation along with allowing for multiple images to be taken of a single specimen in a single view. The previous custom-designed C++ image capture software was modified to incorporate image normalization and control a motorized stage [67]. Once placed randomly on a black opaque slide, specimens are localized, and multiple images of each specimen are captured automatically [67]. To examine the effects of illumination, 18 images were taken of each specimen with the stage rotated in 20° increments and the light direction coming from a fixed 270° azimuth and 30° zenith. The eighteen images are considered to be frames of a digital video of the specimen with changing illumination. This is why the digitization method is considered video capture.

Once the 18 images are captured, an improved computation algorithms component is used. Instead of using single image pairs for correlation coefficient, i.e., one image per specimen, we switch to using the median score of the correlation coefficient between all 18 image pairs [67]. A switch to Agglomerative Hierarchical Clustering (AHC) was made with improved results, as opposed to the previous Maximal Clique Clustering (MCC). AHC performed significantly faster than MCC and obtained slightly better CGR and IGR percentages, and so AHC was adopted into the new system.

In the direct identifications component, a minor change was made to the visualization method of the digital specimens. In the first prototype, the Wiki only provided one image, or at most two images, of the specimen. With CASSIE 2 we are able to provide a more informative digital specimen representation. From the 18 images, videos of a rotating light source were added to the Wiki using JavaScript to cycle through the images like video frames. Once placed in sequence, the images give the illusion of having a fixed specimen with the light source moving around it.

Along with all the changes to the first prototype, research was also conducted into better digital representations. As the goal for this prototype was to create more accurate digital representations, preliminary research into anaglyph (3D) representations was conducted to incorporate depth information. To form anaglyphs, the shape of the specimens had to be estimated using the 18 images. All images were normalized as in the CASSIE 1 prototype and it was discovered that images did not align properly due to the illumination differences. To improve alignment, photometric stereo was used to align the 18 images [67]. Once the images were properly aligned, photometric stereo and depth map estimation were used to create a model of the specimen. Anaglyph images were then rendered using the model. At this stage, anaglyphs were displayed as videos for testing but were not incorporated into the main system.

After the requirements refinement and prototype modifications, we waited for a

specialist to identify the new dataset as promised, for testing and validation. However, the specialist was unable to identify the full 1 000 specimens or a subset thereof in a timely fashion, and there was insufficient time to find a replacement. The difficulty obtaining identifications from specialists or experts was the major motivating factor for the development of the third prototype.

2.1.4 Third Prototype (Microfossil Quest)

As ground truth results are crucial for any research, the difficulty obtaining identifications from experts led to the development of the third prototype. Taxonomic identification is difficult for computers to do and time consuming for experts. We take the strengths of computers and humans and combine both in the Microfossil Quest. For the third prototype, we engineer crowdsourcing into microfossil identification. Crowdsourcing research has shown that engagement of ordinary people to solve tasks is a modern approach to tackling difficult problems. In particular, human-based computation is a new area gaining popularity for solving AI-complete problems. We incorporate human participation using citizen cyberscience to obtain specimen identifications. This change led to the examination of the computer-aided approach that was used, and how the system has now changed to be considered a human-based computation approach supported by citizen cyberscience.

Unlike previous versions, considered computer-aided approaches to identification, the computer-aided term does not adequately describe the new system. Previously, it was assumed that a few humans, namely experts, would be entering identifications and using the system to help obtain identifications for later use. As the system would be helping specific experts or specialists who provide and use the identifications, the original system was seen as aiding the original expert by reducing his workload.

The new system centers around human-based computation because the individuals providing the identifications are not necessarily the same as those that will use the identifications. Although the end goal of the system is still to identify a large quantity of specimens quickly and with high quality results, the need to increase user effort, though not necessarily expert, is a new direction. This type of crowdsourcing system, where humans aid the computer, is considered human-based computation. The system also uses citizen cyberscience because of how user input is obtained. Identifications are obtained online to help research, and the online website will educate the public about microfossils and microfossil research. Research focus, public education, and online identification is why the human interaction component utilizes citizen cyberscience.

With the change from a computer-aided approach to human-based computation using citizen cyberscience, significant changes to system behaviour and focus were necessary. The significance of these changes led to the system as a whole being renamed the Microfossil Quest, which is also the name of the third prototype, or Quest for short. The main goal for the Microfossil Quest is to obtain a large database of identified specimens that can support evolutionary prototyping towards a fully-automated microfossil identification system. The top-down design of the system including the requirements, implementation, and testing of the prototype is described in Sections 2.2 to 2.4.

2.2 Requirements Refinement

Many different considerations were encountered while designing the Microfossil Quest prototype. The purpose of this prototype is to obtain identifications quickly for a large dataset. For this reason, the clustering techniques developed in previous prototypes were further developed and refined to suit the crowdsourcing approach being taken. To incorporate human-based computation and citizen cyberscience, we considered best practices as determined through the review of other citizen science, citizen cyberscience, and human-based computation systems.

The requirements for the system are broken into application-specific and approach-specific requirements. Application-specific requirements were placed on the system as a consequence of experience with the CASSIE prototypes. These requirements are explained in Section 2.2.1. Approach specific requirements centered on key features of any system using a crowdsourcing method, and are described in Section 2.2.2.

2.2.1 Application Specific

Application-specific requirements for the Microfossil Quest focused partly on transferring and improving from the CASSIE prototypes. One transition requirement affecting system design is ensuring backward compatibility. Ensuring all data obtained in previous prototypes can be transferred, viewed, and used in the new system played an important role in the design of the new system.

The second application-specific requirement is obtained from the goal for the end prototype to obtain a high quantity of identifications quickly while maintaining identification quality. To meet this end goal, the Microfossil Quest must be designed to cluster as many individual specimens as possible in a short amount of time without compromising identification reliability. Unlike HiRise Clickworkers [68], Star-

dust@home [50], Galaxy Zoo [55], and all the other spin off projects from Galaxy Zoo, our system incorporates back-end computation to improve the impact of a single user identification and allow us to quickly identify the dataset. Back-end processes evolving from CASSIE 2 include clustering to ensure a high quantity of identification (thoroughness), identification propagation to maintain identification quality (reliability), and prioritization to obtain identifications quickly (throughput).

2.2.2 Approach Specific

Approach-specific requirements focus on the necessary changes to go from the CASSIE prototypes, which are both computer-aided approaches, to the Microfossil Quest, a crowdsourcing approach. Crowdsourcing-specific requirements for the system are determined by examining previous work on citizen science. Citizen science approaches are relevant to our crowdsourcing application because we use citizen cyberscience to engage with volunteers. Best practices of citizen science projects are described in papers on citizen science frameworks. It should be noted that, while some citizen science frameworks focus on physical data collection, depending on the application and design considerations, other frameworks place more focus in different areas. Key factors that should be incorporated into citizen science projects include [69–71]: expert participation; citizen calibration, training, and motivation; data verification; and restriction of malicious users, which is of particular concern for citizen cyberscience.

As expected with most citizen science projects, some expert participation is required. Expert involvement ensures research protocol is followed, valid data is obtained, and citizens are trained. It is the interaction with real researchers that enables citizens to learn more both about the field, and about the scientific process. In our case, micropaleontology experts will be asked to help ensure the validity of our tutorial, and to provide identifications that can also be used for citizen calibration.

Citizen calibration is completed by comparing identifications provided by citizens to the identifications given by experts. Volunteers are separated into trained volunteers, such as experts and specialists with formal training in microfossil identification, and untrained volunteers, such as novices and citizens with no required formal training. Novices would advance to citizens depending on their level of participation and accuracy as determined by comparisons to trained volunteer results. Using the identifications obtained from trained volunteers for calibration, we can decide if or when to upgrade untrained volunteer status. By including the experts into the system, we allow their identifications to create calibration specimens that are no different from other specimens making it more difficult for malicious users to predict calibration

specimens.

Before volunteers are able to provide any usable identifications, they must first be trained at least informally. For untrained volunteers, a tutorial was developed for the Microfossil Quest. This tutorial describes the website and microfossil features in simple terms allowing the average user to understand and interpret the descriptions. Images of the features are also provided to assist volunteers. The tutorial focuses on features that are most commonly used to distinguish specimen taxons. Trained volunteers will not need to go through the full tutorial, but should view the sections describing how to use the website.

Citizen motivation is the last indicated priority for all citizen science projects. Citizen motivation can be done by various methods. A scorecard, or user status, all appear to work effectively as motivating factors provided mechanisms are in place to identify and isolate malicious users. Herbaria@home found volunteers liked to "compete over the number of sheets they can complete" [45].

Closely tied to citizen motivation is also data verification. It should be noted that any methods used to motivate users may also lead to possible reasons for malicious data being entered into the system. Through the examination of several crowdsourcing approaches, it was seen that accounting for malicious users or incorrect data is an important consideration for citizen cyberscience. An example of this was seen in Stardust@home where some volunteers attempted to cheat by "flipping through as many images as possible to rise to the top of a scorecard put in place as an incentive" [49].

It is best to assume that a minority of participants are malicious users when designing citizen cyberscience projects. In Galaxy Zoo, 36 users appeared to record random identifications possibly due to an automated process or browser issues. These malicious users were in the minority, at 0.05% of the total participants, and incorporating a system that is able to both detect and isolate these users is important to ensure the validity of the dataset [72]. Looking at the Galaxy Zoo study, it was mentioned that some users might intentionally give incorrect identifications. However, incorrect identifications are minimized by restricting a user to a limited menu. From the well documented experience of the Galaxy Zoo project, volunteers overall are genuinely willing to help. However, steps should be taken to minimize the impact of the few malicious users on the final results.

As indicated by the literature, there are six key factors that all crowdsourcing projects need to consider: expert participation, citizen calibration, citizen training, citizen motivation, data verification, and restraining malicious users. With these approach-specific requirements in mind, as well as the application-specific ones, we move to the prototype modification stage of the evolutionary-prototyping life cycle.

2.3 Prototype Modification

Due to the major change from a computer-aided to a crowdsourcing approach, many changes were necessary to modify CASSIE 2 into the Quest prototype. For the prototype modifications, improvements have been made to the back-end processing along with incorporating a new front-end website. A component diagram of the new system is shown in Figure 2.5. Many components dating from the CASSIE 2 system required modification, and in some cases complete revision. A description of the changes made to the system components are described in Section 2.3.1. The programming languages and software architecture used by the components are explained in Section 2.3.2.

2.3.1 System Components

The Microfossil Quest prototype is a significant change from CASSIE 2, the previous one. As can be seen looking at Figure 2.5, specimen acquisition remains the same. The remaining components—specimen dissemination, human interaction, and computation algorithms—underwent minor to significant changes.

The specimen dissemination component was initially developed in CASSIE 2 to create anaglyph videos that were not incorporated into the main system. For the Quest prototype, Harrison further improved these digital representations to incorporate anaglyph and illumination rendering combined together.

A major decision was taken to redo the human interaction component used to interact with humans. Previously, the Wiki website was designed to be seen and used by a limited number of users, as a component of the CASSIE prototypes, to obtain expert identifications. In the Microfossil Quest, the website is the human interaction component of a crowdsourcing system. Many users are expected to use this website, and with the expected longevity of the system, a new design for the website was needed to make it easier to maintain and modify. Designing a new website incorporating crowdsourcing involved a complete redevelopment of the human interaction component of the system. This new website is the Microfossil Quest front-end seen by citizen cyberscience users and includes the database, website-database interaction, and anaglyph with illumination visualization. A detailed description of the website is given in Chapter 3.

The design of the algorithms, or the back-end of the system, underwent significant changes. Previously, in CASSIE 2, the computation algorithms component occurred immediately after specimens were obtained. Once clustered templates were chosen automatically, identifications were required for those templates. In the new approach, we do not force users to identify particular templates. Instead, we suggest speci-

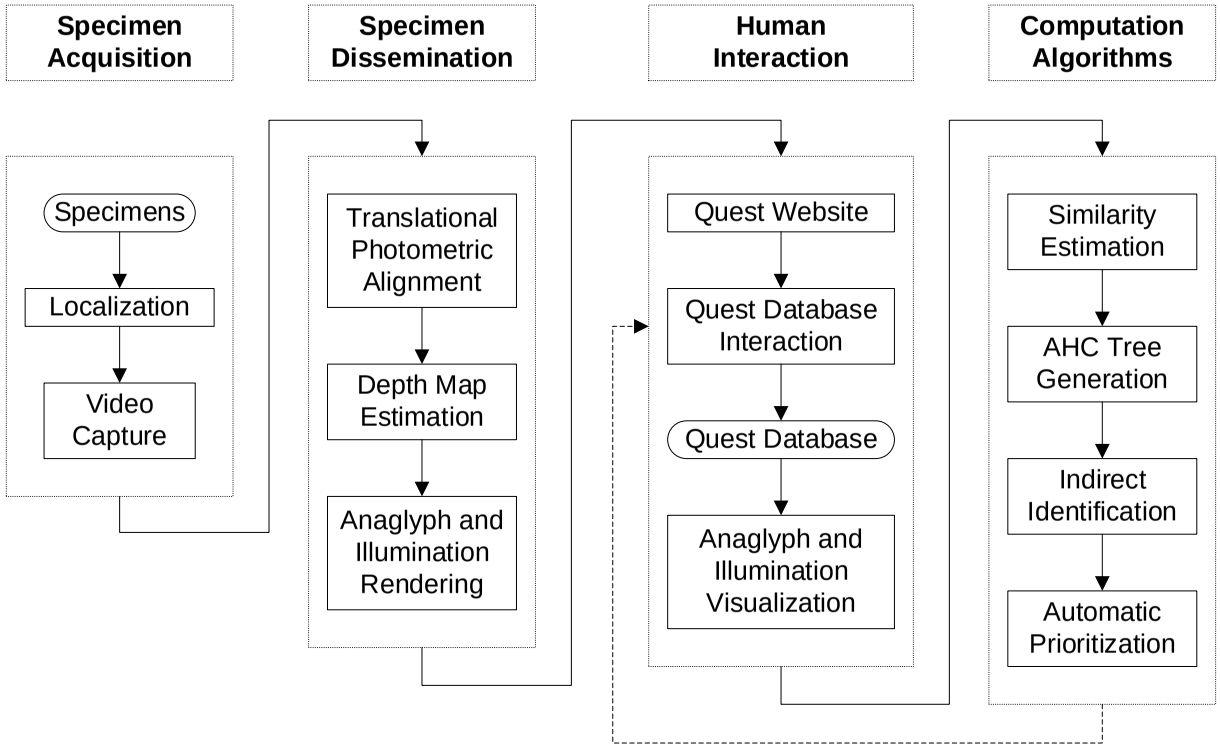


Figure 2.5: The system components and data flow for the third prototype. Specimen dissemination has been improved and incorporated into the main system with direct identifications being completely redesigned. The computation algorithms component was modified and incorporates indirect identification and automatic prioritization algorithms.

mens to be identified but users are allowed the freedom to identify any specimen in the dataset. The new computation algorithms occur mainly after identifications are obtained as the cluster results are regularly updated after volunteers provide identifications, which impact the suggestions. After similarity computation, using image correlation coefficient for simplicity, a new clustering method employs AHC to generate trees for use to improve identification thoroughness.

Indirect identification and prioritization algorithms were developed to increase identification impact. It can be expected that individual users will spend varying, and sometimes very limited, amounts of time identifying specimens. Participation may also vary at detectable levels in relation to publicity of the project, as was discovered during the Galaxy Zoo study [54]. Due to the unpredictable and irregular nature of volunteer participation, the clustering algorithm was created for thoroughness, identification propagation for reliability, and specimen prioritization for throughput. All three algorithms are described in Chapter 4.

The specimen dissemination, human interaction, and computation algorithms components all incorporate the necessary functionality for the Microfossil Quest. In order to program these components, a combination of different languages and architectures were chosen depending on those best suited to each component.

2.3.2 Languages and Architectures

The Quest prototype combines many different programming languages and software architectures that were used for various reasons. A summary of the implementation is given in Table 2.1. Components indicated in this table refer to the components shown in Figure 2.5. The languages used were chosen based on implementation time and function, while architecture choices focus on readability and maintainability of the system software.

Languages

The languages chosen for the website and background programming of the Quest prototype were important as it influenced the amount of time needed to implement the system. For the specimen acquisition component, the C++ language choice was decided in previous prototypes and left unchanged. Many of the language decisions for the modified system were influenced by an available framework, desired functionality, and compatibility with already chosen languages.

A framework for the Quest system was provided by Ranaweera, who was the developer of the Microfossil Wiki [66]. This framework used basic HTML, PHP, and

Table 2.1: The Microfossil Quest comprises human and computer intelligence, which are built from different components. For each component, the programming languages and software architectures are given.

System Intelligence	Component	Languages	Architecture
computer	specimen acquisition	C++	pipe-and-filter, client/server
	specimen dissemination	MATLAB	pipe-and-filter
	computation algorithms	MATLAB	pipe-and-filter, client/server
human	human interaction	HTML, PHP, CSS, JavaScript, AJax, Java, MySQL	pipe-and-filter, peer-to-peer, model-view-controller

CSS. Adopting this framework allowed for reduced time in setting up the model-view-controller database and preliminary database interaction. Ruby on Rails was considered, but learning this for the prototype was not desirable due to time constraints. From this given framework, other languages were included to meet desired website functionality in the limited time frame.

Within the website, certain desired functionality influenced the decision to use JavaScript, Ajax, and Java. JavaScript was used to program the menus, specifically to open new windows that interact with the main browser and that have dependent menu options. The website also incorporates Ajax and posting functionality from the Yahoo UI library to transfer input/output information and reload specified website areas instead of having to refresh the full webpage, which would take more time. Java was used in the human interaction component to perform the non-anaglyph, anaglyph, and illumination visualization. By using Java and a Java applet, the processing required to generate a specimen image from the model is done on the client computer, making it faster and reducing the amount of data transfer needed. To reduce transfer time, the data used by the Java applet is stored in a MATLAB-generated binary file, which holds the image size and image data—equivalent to 9 greyscale images (3 and 6 for non-anaglyph and anaglyph representations, respectively).

The remaining requirements for the system included database and back-end processing languages that were decided based on compatibility with previously chosen languages. MySQL is used for the database and MATLAB for the back-end algo-

rithms. The database was implemented using MySQL because of its high compatibility with websites and PHP. MATLAB is a powerful environment for data-driven programming, which works well with images and matrices, so we used it to develop the back-end algorithms. After determining that MATLAB was also able to interact with the MySQL database, the identification and prioritization calculations were kept in MATLAB to avoid unnecessary reprogramming.

Many of the language choices were determined from available framework, desired functionality, and compatibility. These choices led to a complex system that was prototyped relatively quickly, while maintaining desired system functionality. Using these languages, the architectures for system components were also considered to ensure portability, conceivability, and modifiability.

Architectures

Program code can be organized using several different architectures depending on the type of access required. The various architectures used in the Quest prototype are defined before we justify the use of the most common architectures, namely pipe-and-filter, peer-to-peer, client/server, and model-view-controller.

Many architectures used in the Quest prototype are common architectures defined in software engineering. The Quest prototype uses the pipe-and-filter, peer-to-peer, client/server, and model-view-controller architectures for decomposing the system [73]. Pipe-and-filter has subsystems process data from the input stream and returns results as an output stream. Peer-to-peer architectures have subsystems that are able to function as both a client and a server. The client/server architecture has a distinct client asking for information, and a distinct server providing information. Model-view-controller separates code into models that store information, views where information is displayed or represented, and a controller that performs any processing on data.

The pipe-and-filter design is implemented throughout the full system both in the intra-component architecture, within a component, and inter-component architecture, between components. Each component that is called uses combinations of functions, or filters, to perform calculations, and itself is a filter with inputs and outputs. Pipes and filters were chosen for modularity, making alterations and improvements easier, as well as allowing for the reuse of common functions. An advantage of this approach is the ability to improve individual filters to obtain more reliable and accurate results in a shorter time frame. Should future algorithms improve results in one area, the system can be modified relatively easily. The modularity also makes it easier to add or remove functionality as each process acts independently of others.

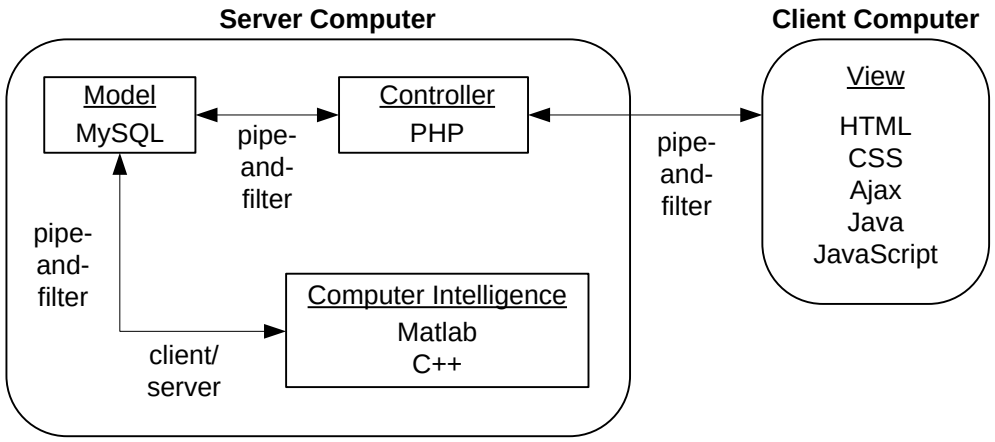


Figure 2.6: A graphical depiction of the system code organization, execution location, inter and intra-component interaction, and languages.

Peer-to-peer architecture is utilized in human interaction because it acts as a client or a server to different components. Human interaction acts as a client to the specimen acquisition and computation algorithms components. In addition to this, it also acts as a server to web browsers. Because the human interaction component behaves as a server and a client, it uses the peer-to-peer architecture. The structure of the human interaction architecture is depicted in Figure 2.6.

The Microfossil Quest uses client/server architecture in the specimen acquisition and computation algorithms components. The difference between these components and the human interaction component is that their architecture does not change. The specimen acquisition and computation algorithms always behave as servers. For this reason these two components use the client/server inter-component architecture.

A model-view-controller setup is used in the intra-component architecture for the Microfossil Quest human interaction component. This architecture is mainly employed in the website where the views are the browser sites, the models are the database table specifications, and the controller performs processing needed to determine what views are shown and what items are displayed in the view. The model-view-controller architecture helps separate the programs and functions of the website, making future development easier. Figure 2.6 shows how the system files are organized, where code is executed, how the components interact, and where languages are used. All MySQL database table specifications and functions are stored in the model folder and execute on the server. View folders hold all the basic PHP, HTML, and JavaScript files, which describe the web page layout and text and which execute on the client. The controller folder holds any files written in PHP that select and control the content displayed. The model-view-controller architecture provides modularity,

making it easier to make modifications to various aspects of the website. For example, the views seen by users is expected to go through many changes in order to find an appealing website for returning and new volunteers. However, the database itself is fairly static. Once launched, the database tables will not undergo significant change and should not be affected by any alterations to views or controllers.

The architectures used in the Quest prototype were defined and a justification was given for the use of pipe-and-filter, peer-to-peer, client/server, and model-view-controller architectures. These architectures support the portability, conceivability, modifiability, and long-term life cycle of the system by making it easier to maintain and expand. All the implementation decisions from the component breakdown, languages, and architectures were chosen to implement the new system requirements while supporting a long-term and evolving system. To verify functionality, the Quest prototype was then tested and validated.

2.4 Testing and Validation

At the testing and validation stage of the Quest prototype, focus was placed on testing the components of the system as opposed to the system as a whole. Because there have been major modifications to CASSIE 2, the priority was to validate the expected functionality of the modified components: specimen dissemination, human interaction, and computation algorithms.

The specimen dissemination component was developed to incorporate better digital representations. Preliminary testing for this component was completed in the CASSIE 2 prototype through the creation of anaglyph videos. Further testing was conducted in the Quest prototype to ensure rendered illumination for a specimen matched with obtained images where azimuth and zenith angle are set at 270° and 30°, respectively. Once the algorithms were verified, the setup was modified to enable illumination, non-anaglyph, and anaglyph control using a Java applet called Virtual Reflected-Light Microscopy (VRLM). The correct functionality of the VRLM applet was verified through testing the user interface, image display, illumination rendering, and anaglyph rendering. Once the functionality was verified, minor testing and tweaks were conducted to ensure the rendered anaglyphs closely matched the physical depth information that can be seen through an optical microscope, ensuring a realistic representation. Using the applet, VRLM digital representations are viewable at `www.ece.ualberta.ca/~imagesci/vrlm`.

The human interaction component contains the website and database functionality. To verify the correct functionality of the website, navigation of the website was

tested through the menu and links. For both testing purposes and backward compatibility, the Wiki dataset was loaded into the Quest database. Using this loaded dataset, the functionality of the search and caption menus were verified. The message forum was also tested by creating an example topic thread with some replies. Lastly, VRLM functionality was tested within the Microfossil Quest website. By using all these different aspects of the website, we have verified the correct functionality of the Microfossil Quest website.

Computational algorithms behind the Microfossil Quest were tested more in depth to ensure expected functionality. Through development, the expected functionality of the unsupervised, supervised, and dynamic learning algorithms were examined with a small subset of specimens in the dataset. Once the correct behaviour of each algorithm was verified, the next step was testing the expected impact each algorithm had on the dataset. From our expectations, the unsupervised algorithm ensures thoroughness of identifications in the dataset, the supervised algorithm ensures reliable identifications are propagated, while the dynamic learning algorithm ensures a predictable and reliable dataset identification, leveraging the advantages of the supervised algorithm, while ensuring less time required to fully identify the dataset. More details on these testing procedures and results are found in Chapter 4.

Testing and validation of the Microfossil Quest prototype was conducted individually for the specimen dissemination, human interaction, and computation algorithm components. With behaviour of each component verified separately, we can confidently state that the Microfossil Quest prototype satisfies our requirements and successfully translates from a computer-aided to a crowdsourcing approach.

2.5 Conclusion

Previous research into a fully-automated microfossil identification system has met with limited success. In the development of a fully-automated, economical, and reliable system we use a different approach to development. Instead of conceiving and implementing a full system at one time, we use an evolutionary prototyping approach that is ideal for time constrained, crowdsourcing, and research projects. During system development, previous prototypes were eludicated and the latest Microfossil Quest prototype requirements, modification, and testing stages were described.

Prototypes naturally evolved through the discovered deficiencies and needs determined after testing and analysis of results. The Microfossil Quest was designed to obtain specimen identifications after discovering a large bottleneck when testing CASSIE 2. Improving methods to obtain identified specimens will assist with research

requiring ground truth results to train, test, and validate system performance.

The requirements for the Microfossil Quest is different from previous prototypes. As opposed to focusing on a computer-aided approach to microfossil identification, we now use a crowdsourcing approach. This change necessitated application and approach-specific requirements during development.

CASSIE 2 was heavily modified and initial tests were conducted on modified system components. It was decided to create a new website front-end, described in Chapter 3, and new back-end processing algorithms, explained in Chapter 4. The basic website functionality has been verified, along with algorithm testing and validation, which is further described in Chapter 4.

Prototype evolution, and the latest requirements, modification, and testing descriptions for the Microfossil Quest prototype were outlined. As this system evolves and grows, we hope to reduce the amount of reliance on citizen cyberscience and humans to reach a fully-automated microfossil identification system.

Chapter 3

Human Interaction

Identifying specimens is the most difficult aspect of automatic identification. In the Microfossil Quest system, we determine identifications with the help of a human-interaction front-end. The website front-end is an example of citizen cyberscience, where users can participate and learn more about microfossil research. The layout, design, and content is important as potential volunteers use the site to provide identifications and learn about the project. Due to the importance of the impression made by the website, previous crowdsourcing projects were reviewed to determine how digital representations are displayed, how identifications are obtained, the content to include, and the functionality to include.

Humans place great importance on sight and, because of this, how digital representations are displayed is important. Having a visually appealing and interactive website is important to ensure volunteers remain interested in the project. Stardust@home, ESP Game, and Zooniverse projects all use different methods to achieve this. Stardust@home uses JavaScript and HTML to create a 'virtual microscope' [51]. This interactive microscope allows users to scroll through a stack of 43 images to change the zoom level of displayed aerogel tiles. The ESP Game provides an appealing and interactive website through a Java applet [47]. Depending on specifications, different programming languages and software architectures can be used as in the collection of crowdsourcing projects under Zooniverse. Zooniverse projects use a variety of methods to display data—the Hubble project uses JavaScript, Ajax, and HTML, while Understanding Mergers conducts human interaction through a Java applet [64]. All these methods of displaying and interacting with the system were developed to ensure a visually appealing and easy-to-use interface, which are important considerations when developing crowdsourcing projects.

How information is obtained from volunteers plays an important role in the quality of the data obtained. HiRISE Clickworkers and Galaxy Zoo obtain data through

the use of big buttons with images to help volunteers distinguish between different possible identifications [55, 68]. An issue with using big buttons is the possibility of biasing results if images are chosen inappropriately. In terms of taxonomy, using images would not only bias results, but the space required to create buttons for all the possible species would not be feasible.

In contrast to buttons, Verbosity obtains information through open text boxes. Players are asked to fill in blanks located in six sentences to describe a word that must be guessed by a second player [60]. It was seen that the freedom for users to enter their own information allowed the opportunity for malicious users to take advantage of the system. Through analysis of the system, it was seen that "people cheat at Verbosity" [60]. The sentences were meant to direct players to fill in the text boxes using responses that would provide specific relationships with the guess word. However, "the describer often ignores the relation and says what they mean" [60]. Because players are given an open text box to provide their hints, some players cheat by typing a word very similar to the one to be guessed but not appropriate within the sentence. In other cases, players would type the word itself across several hints. Having more control over user input would prevent malicious volunteers entering undesirable input, and normal volunteers from unintentionally entering incorrect input that would have a high impact on the further processing done on the dataset.

One input method used to minimize mistaken data is seen in Herbaria@home, which uses a combination of text boxes and menus [45]. Herbaria@home allows users to manually enter information using open text boxes. However, for some fields, a drop-down list will also appear giving users the option to select from a list of known options. This approach prevents sincere users from unintentionally entering incorrect information. Unfortunately, it does not restrict malicious users.

Aside from the visual appearance and interaction of the website, the content of the website must also be determined. The website is designed to support citizen cyberscience, which incorporates educational material. All citizen cyberscience projects educate volunteers, mainly through a tutorial. The information about how the system itself, or part of the system, functions is not always provided on the website for volunteers to view before or while participating. However, in the case of Stardust@home, software information is included through a description of how the 43 virtual microscope images are obtained [51].

While reviewing the literature on crowdsourcing projects, it can be seen that forums are also important content. Foldit, Galaxy Zoo, and Stardust@home all incorporate a forum, where citizens are able to interact with each other and developers. Cooper *et al.* [40] state that the popular Foldit game uses many different reward

structures to encourage and prolong volunteer engagement. These include chats and forums where social praise encourages users. Stardust@home also has active forums where "participants have extensive discussions, and have named themselves *dusters*" [51]. Galaxy Zoo has an active forum for volunteers to interact. These forums assist projects by forming a community of volunteers, allowing them to interact with other members with similar interests, and encouraging volunteers to return to the website.

Website functionality can also have an impact on how volunteers react to the system. For most projects, how objects are chosen for volunteers is not indicated. The exception to this is Galaxy Zoo, where galaxies are shown randomly [54], and Herbaria@home, where herbaria sheets can be identified randomly or filtered by herbaria project, collection, or genus [45]. Normally, Galaxy Zoo information is easily distinguished visually by humans, so randomly showing galaxies would not have a large impact on volunteer experience. This is unlike Herbaria@home, where many more details are requested from the volunteers. In this case, allowing the volunteers some control over the sheets they identify helps to ensure volunteers are not overwhelmed and feel they are forced to provide identifications in situations where they are not comfortable. It should be noted, in this example, that the amount of control volunteers have is still limited as they are only able to filter by herbaria project, collection, or genus. The actual sheets shown to users are still chosen randomly from the filtered results.

From our review of the literature, it was decided key features to include were an interactive digital representation, control over user input, more software descriptions, a member forum, and volunteer control over specimens to identify. Once the content was established, the website had to be designed. The basic structure of all the pages in the Microfossil Quest website includes a header, body, and footer. The header contains the project title and navigation menu, while the body changes and the footer remains empty at present. Table 3.1 shows the navigation menu, which depicts the structure of both the website and this chapter. The menu has several different headings organized from most specific, on the left, to most general, on the right. Each subheading in the menu links to a different web page on the website, and also indicates sections in this chapter. The text in Sections 3.1 to 3.5 is a copy of the text displayed on the website, except when shown in italics.

Table 3.1: Navigation menu for the Microfossil Quest website. The first row in the table are the menu headings. When the headings are clicked, the submenus, linking to more information, are revealed.

Home	About	Tutorial	System	Background
	Overview	Overview	Overview	Overview
	FAQ	Website	Users	Microfossils
	Forum	Microfossils	Acquisition	Crowdsourcing
		Shell Textures	Human Intelligence	References
		Chambers	Computer Intelligence	
		Apertures	Knowledge Base	
		View Sides		

3.1 Home

The main purpose of the Microfossil Quest front-end is for volunteers to provide taxon identifications. To place emphasis on providing identifications and to make it easier for returning participants, the search and identification page is displayed as soon as the website loads. This home page is where volunteers will be able to search the database and provide identifications by updating specimen captions. A screen shot of the home page is shown in Figure 3.1. The body of the home page is divided into three separate areas: search, digital representation, and caption.

The top area on the home page is the search area. Volunteers are able to begin identifications on the default list of specimens, or search for a specific subset of specimens to identify. For example, if a volunteer is unable to provide taxon identifications, but is comfortable providing view side information, they can search for specimens with unknown views. This allows volunteers to select a subset of the database focusing on specimens lacking data specific fields.

*Underneath the search section is the digital representation area. This is where two possible digital representations are available. Volunteers are able to view the images taken of the specimen from the microscope or another more detailed representation. To create more realistic digital representations, work was conducted as part of the Computer-Aided System for Specimen Identification and Examination (CASSIE)*2 *prototype to generate anaglyph representations that provide depth and illumination information. We determined that allowing users to control illumination direction was more engaging and informative compared to providing a video showing different illumination conditions. As part of the Quest prototype, this algorithm was expanded to render images dynamically through a Java applet called VRLM.*

The VRLM applet renders and displays images while giving users control over il-

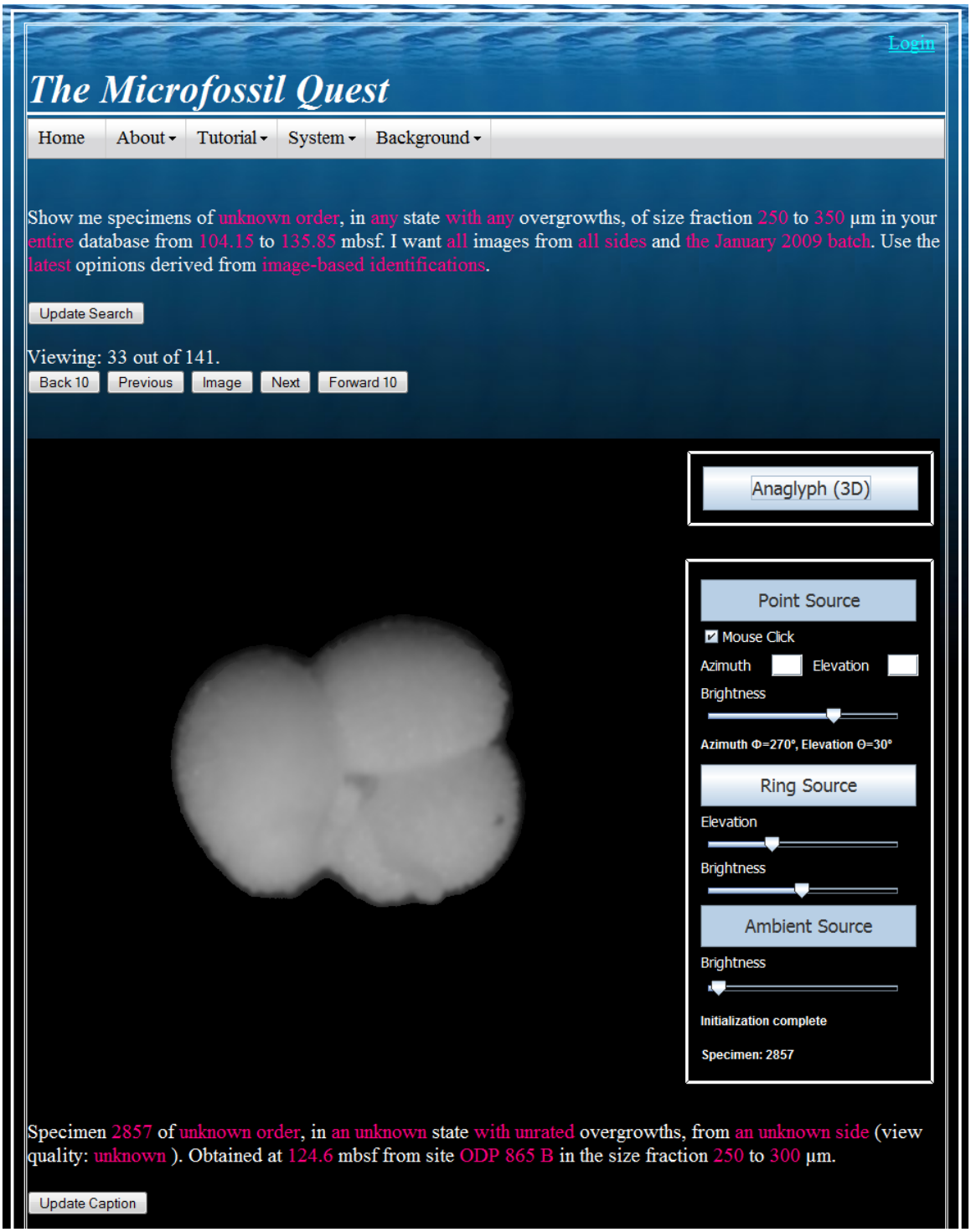


Figure 3.1: The home page showing the VRLM applet view for one specimen in the database.

lumination source, brightness, and direction. In this way, it provides a more realistic microscope environment. It allows all specimens to be viewed as non-anaglyphs or anaglyphs, which display depth when seen through red-cyan glasses. A visual snapshot of the VRLM applet is seen in the digital representation area of Figure 3.1. In comparison to other projects, the applet provides a more complex environment than Stardust@home's 'virtual microscope'. Unlike Stardust@home, our virtual microscope uses a surface model to generate images, as opposed to displaying existing photos. VRLM was also designed to minimize the amount of data transfer required. All the information required for illumination control, depth perception, non-anaglyphs, and anaglyphs is stored in the space equivalent to 9 *greyscale images, which is significantly less than the* 43 *colour images used in Stardust@home. More details on the materials and methods used to create VRLM representations is described by Harrison* et al. *[74].*

The last area on the home page, the caption area, is found near the bottom. When a user is at a specimen, they will view a caption describing any information belonging to that specimen. If an identification has not been made, the caption will indicate it. Volunteers are able to provide direct identifications by updating this caption on the home page. Details on user interaction with the caption is given in the tutorial, i.e., in Section 3.3.

In summary the body of the home page is separated into the search, digital representation, and caption sections. Users are able to filter the specimens they wish to identify, view the digital representations of specimens, and modify captions to provide direct identifications. The home page contains the main functionality for the website, and provides the user interface connecting to the database.

3.2 About

The about section is an introduction for new users. The purpose of this page is to grab the attention of potential volunteers, to encourage them to participate, and to invite them to learn more about the Microfossil Quest. The about section includes an introduction for new volunteers, answers to frequently asked questions as described in Section 3.2.1, and a forum, as described in Section 3.2.2.

Have you ever heard of Foraminifera? How about Dinoflagellates or Radiolaria? These are all different orders for microorganisms that can be found around the world. You may not know what they are, you may never even realize that you have seen them, but these small animals are important to researchers. Some microorganisms, like the three that are mentioned, are special because they can be found fossilized as sediment on the ocean floor. Micropaleontologists study these microfossils using

various microscopes in order to see the details found in these microfossils. To most people without any equipment, we may mistake these fossils for grains of sand. In fact, the limestone used to make the great pyramids of Egypt were partially made of foraminifera [75, 76], and chalk is made of various microfossils, including foraminifera and coccoliths [15, 77].

The Microfossil Quest is dedicated to assisting micropaleontologists. At this website, ordinary people like you can help scientists by looking and identifying pictures and anaglyph (3D) models of microfossils. Obtaining these identifications is essential to support research conducted using the microfossils. The amount of information you provide is up to you. For those who want to take things slow, you can try to look for a few easy-to-find features in microfossils. Those of you who are more interested are encouraged to learn and provide Linnaean taxon (order, genus, and species) identifications if possible. If you are still not sure, feel free to take a quick look around our site, at our tutorial, or at the microfossils stored in our database. If this is something that might interest you then try it out and join the Quest. Help researchers study prehistory through microfossils.

3.2.1 FAQ

The Frequently Asked Questions (FAQ) section is used to answer questions that are asked repeatedly by users. This provides an area where users can easily find answers to questions that we did not anticipate, or have not addressed in other areas of the website. This also prevents users from posting the same question multiple times in the forum. At this stage, only information regarding the VRLM applet is on this page.

Answers to FAQ can be found in this section. This includes common questions sent to the developers, popular questions seen in the message forum, and questions that cannot be answered elsewhere on the website.

How can I get the applet to work? The Java applet is coded using Java 1.6 and requires the Java runtime environment. The Java software for Windows or Linux systems can be obtained from `http://www.java.com/en/`. For Mac OS X, version 10.5.2 or later is required to run Java 1.6, and the default Java version needs to be set. Java 1.6 for Mac may be obtained from the Apple website or using the Software Update application.

3.2.2 Forum

The forum section of the website enables users with an account to provide feedback and comments. This section was included because of the importance of allowing volunteers

to communicate with each other. Having a forum helps create a sense of community among volunteers, and provides a location for members to ask questions, answer questions, discuss the project, discuss interesting observations, and discuss other topics of similar interest. As the Quest front-end is centered around web-based interaction, having such a location for volunteer input and feedback is an important feature, as seen in Foldit, Stardust@home, and Galaxy Zoo.

The current forum implemented on the website allows for basic operations. Volunteers are allowed to create new topics, view topics, and post replies to a specific topic. A screen shot of an example topic thread can be seen in Figure 3.2.

3.3 Tutorial

One of the most important website features in any citizen cyberscience project is the inclusion of a tutorial to train volunteers. Tutorials are needed to educate volunteers and ensure information is gathered correctly. The linear progression of the tutorial develops from features that are easy to distinguish for a novice to features requiring some knowledge of the species for accurate identification. Due to the complexities and the varying knowledge of users, our tutorial takes advantage of the non-linear writing available on the Internet. We have organized the tutorial in such a way that users are able to access topics easily and quickly through the overview and navigation menu. In this overview, the tutorials for each topic may be accessed through links using the topic headings. Users may focus on topics they do not understand or they can review select topics. The text displayed under each tutorial heading on the website is given in Sections 3.3.1 to 3.3.6.

In order to perform identifications we recommend you go through this tutorial to obtain a better understanding of the website and features distinguishing taxonomy. Below is a list of all the topics covered in the tutorial. You can come back to this tutorial at any point. Keep in mind you are able to leave things as unknown if you cannot make an identification.

Website

Introduction on how to provide identifications. Includes demos for how to search the database and edit specimen captions.

Microfossils

Introduction to various microfossils that may be seen and how to recognize different orders.

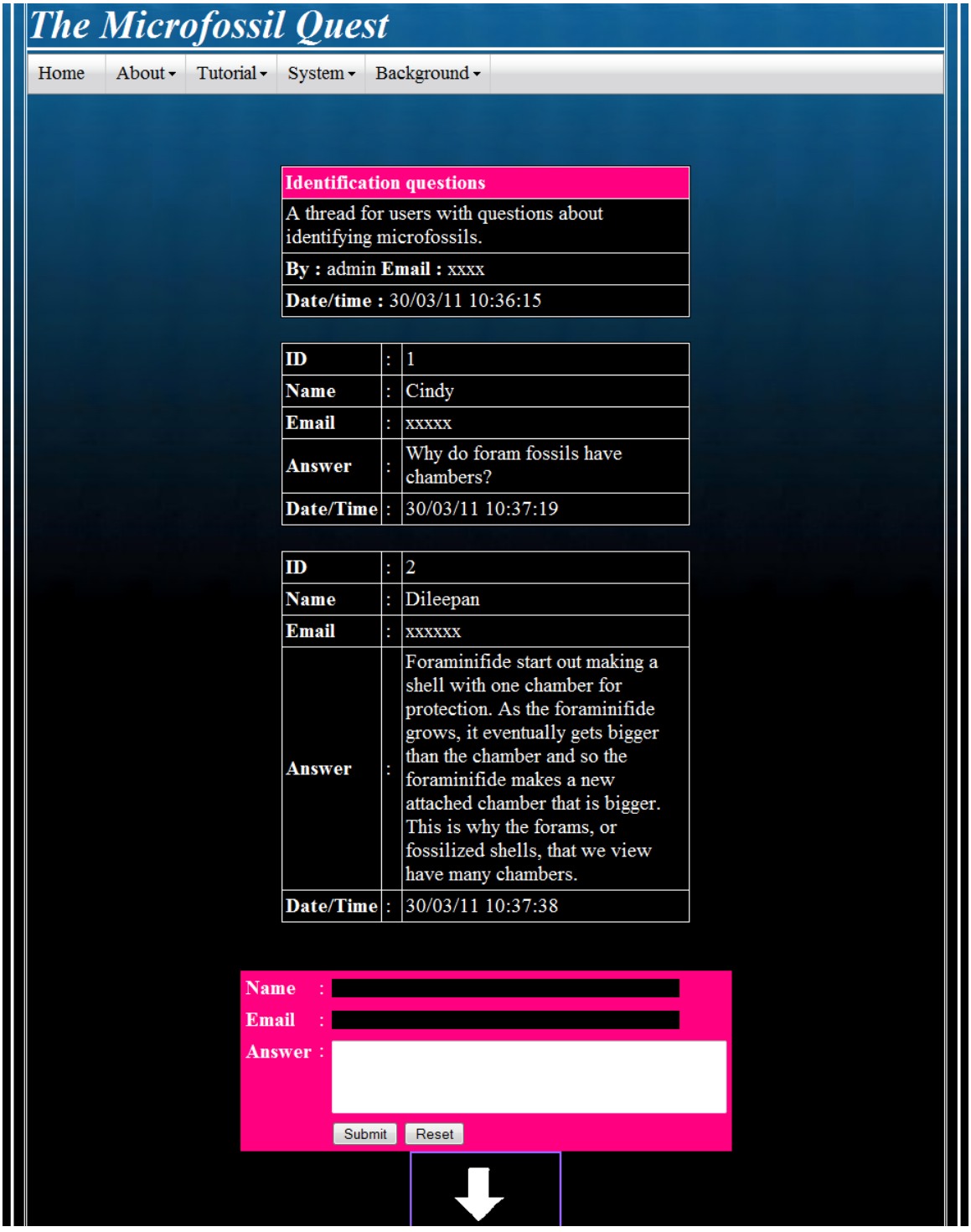


Figure 3.2: Screen shot of a single topic thread in the message forum.

Shell Textures

Description of different textures found on foram shells.

Chambers

Images and names of the various chamber arrangements found in forams.

Apertures

Images and names of the different apertures (or openings) seen in forams.

View Sides

Descriptions of the sides of a foram that may be seen in a view.

3.3.1 Website

All main interaction with the Microfossil Quest is done on the home page. A user can search for microfossils on the website or edit information for a particular microfossil. This splits the website into search and caption sections. The fields that can be specified in the menus are shown before searching and captioning are explained.

Fields

To eliminate any confusion or ambiguity, we define all the fields found in the search and caption menus in Table 3.2.

Searching

The search and caption menus are in a separate pop-up window to simplify the main window. Paragraphs in the main window are used for easy review, while the structured menus make it easier for users to change options. The menus used to obtain identifications are drop down menus to prevent incorrect data from intentionally or unintentionally being entered into the database.

To see a demo of searching, click on update search to open a new window that allows you to specify your options. Once this is done, you can click update which will perform your specified search, update the search paragraph, and close the search menu window. A demo showing the search paragraph and search menu is given. When different options are specified in the search menu, the paragraph will be updated automatically. *An example search paragraph is presented along with the associated search menu in Table 3.3.*

Table 3.2: Explanation of all fields users are allowed to specify in the search and caption menus.

Field	Description
order:	taxonomic order of specimen
genus:	taxonomic genus of specimen
species:	taxonomic species of specimen
brokenness:	broken or unbroken specimen
overgrowths:	extra growths seen on the specimen
quality:	automatically determined quality of the digital representation (good, fair, or poor)[1]
side:	specimen side seen in the digital representation
oblique:	specimen side is seen at an angle (oblique) or not (acute)
identification type:	type(s) of digital representation that was/were seen by the volunteer(s) who identified the specimen
identification view:	view(s) available to volunteer(s) when the specimen was identified
database:	geographical location or sample collection the specimen is associated with, defined by the source the specimen is from
batch:	label given to the group of specimens that are loaded into the Microfossil Quest system together
size fraction:	aperture of sieves used to filter specimens before digital representations are captured
depth range:	meters below sea floor at which the core sample was extracted, commonly abbreviated as mbsf

[1] Calculations used to determine quality are based on sharpness, as described in Appendix B.

Table 3.3: Search menu for "Show me specimens of **benthic foram (Subbotina)**, in **an unknown** state **with or without** overgrowths, of size fraction **minimum** to **maximum** μm in your **ODP 865B** database from **shallowest** to **deepest** mbsf. I want **all** images from **the dorsal (acute) side** and **the Nov 2006(b) batch**. Use the **latest** opinions derived from **image-based identifications**."

Specimen Identification	
order:	benthic foram
genus:	Subbotina
species:	all
brokenness:	unknown
overgrowths:	with or without
Digital Representation	
quality:	all
side:	dorsal
oblique:	no
identification type:	image-based
identification views:	one view
Specimen Information	
database:	ODP 865B
batch:	Nov 2006(b)
size fraction:	minimum to maximum
depth range:	shallowest to deepest

Table 3.4: Search menu for "Specimen **2702** of **planktic foram (Morozovella)**, in **a broken** state **with unknown** overgrowths, from **the dorsal (oblique) side** (view quality: **unknown**). Obtained at **107.15** mbsf from site **ODP 865B** in the size fraction **250** to **300** μm."

Specimen Identification	
order:	planktic foram
genus:	Morozovella
species:	N/A
brokenness:	broken
overgrowths:	unknown

Digital Representation	
side:	dorsal
oblique:	yes

Captioning

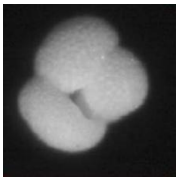
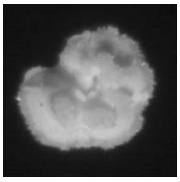
Clicking update caption will open a new window where a new or revised identification may be provided for the current specimen being viewed. Once you finish editing the menu options and click update, the window will stay open so you are able to continue editing the caption for the next fossil. Closing this window or clicking cancel will ignore any edits for the current microfossil, but all previous edits will be stored. A demo showing the update caption functionality is given, any changes to the demo menu will not be stored in the system. *An example caption paragraph and associated caption menu is shown in Table 3.4.*

3.3.2 Microfossils

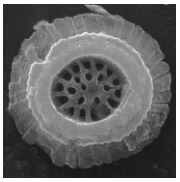
Our definition of a microfossil is any microscopic organism that has been preserved after death. Many different types of microfossils may be found when identifying specimens, it is important to be able to distinguish the microfossils of interest. Table 3.5 shows the various types of microfossils that might be found and describes some distinguishing features of each.

Table 3.5: Microfossil descriptions with example images. Information in the descriptions are from Armstrong and Brasier [15]. All images, except foram images, are taken from Gore [78]. Foram images are taken from our database.

Forams can be found at many different ocean levels and have shells made of calcium carbonate. Planktic forams are usually found near the surface, while benthic forams are found deeper in the water and have a wider variety of shapes.

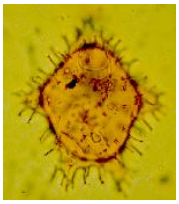
Coccolithophores are formed from coccolith scales, where the most likely ones found are from heterococcoliths. These vary in structure but many are circular or elliptical with radial symmetry. Scales tend to dissolve or disaggregate further down in the ocean sediment and very few coccoliths remain at over 3–4 km deep.



Conodonts are 0.25–2 mm long teeth. These are the dental remains of tube-like jawless worms and are usually found scattered.

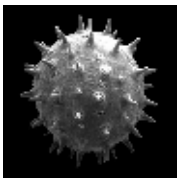

Diatoms have thin and porous shells that dissolve easily and rapidly. Less than 5% end up in sediment on the ocean floor. The main ones that reach the ocean floor are frustules and statospores.


Dinoflagellates average about 20–150 μm in diameter. Dinoflagellates can have many different textures: smooth; grain-like; ridged; and indented. They may also have different shell ornaments: raised crests; short spines; and processes or horns.



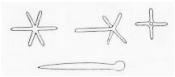
Ostracodes are normally 0.5–3 mm long as adults, but some can be 30 mm long. They usually appear bean-shaped or kidney-shaped. The shells come in two parts, which are connected together with a hinge on the dorsal margin.

Pollen and spores are not as commonly found in ocean cores. Pollen can range from small, simple and spherical to large, bisaccate and intricate. Usually pollen grains have one or two sacs, with the odd grain containing three. Spores can have many different kinds of shapes but are rarely found in ocean sediment. They can have one, two, or three distinct sections that are connected together. Spores are usually symmetrical horizontally or radially. A spore may have many different surface textures: smooth; coated with small grains; covered in mesh-like or fine parallel grooves; warty; and containing rod-like, pointed, or club-shaped projections.



Radiolaria cells are usually 50–200 μm in diameter. They have very porous shells and can have radial or tangential elements or both. Radial elements can be loose spicules, spines, or internal bars that are hollow or solid. Tangential elements are usually a porous shell with a sphere, spindle, or cone shape.

Sponge spicules are the fossilized remains of marine sponge animals. These are usually symmetrical.



3.3.3 Shell Textures

Shell texture is an important feature to consider when identifying microfossils. In the case of forams, the surface texture of the microfossil is used to distinguish between different species. Surfaces may be covered with pores or small spikes, or may be smooth [79].

3.3.4 Chambers

Microfossils, like many things in biology, are identified according to different features. Forams have several main features that can help distinguish between different classes. One of these features is the shape and arrangement of chambers. A chamber is a division of the shell, like a room in a house. These chambers can be arranged in different ways and usually follow a pattern for a particular species. Figure 3.3 lists the common types of chamber arrangements and the names for each.

3.3.5 Apertures

Like chambers, apertures or openings can be used to differentiate forams. When forams are alive, the apertures are used to feed and to excrete waste. Apertures can be in different locations and can have different sizes and numbers depending on species. Common aperture arrangements are shown in Figure 3.4.

3.3.6 View Sides

Currently, we can only see one side of the foram at a time when using digital models or images to identify specimens. Different sides of the specimen can provide different amounts of information to help with identification. Determining what side of a specimen is seen is important when analyzing identification results. It is also important to remember that different sides of the same microfossil can look very different from each other. A description of the various view sides is given in Table 3.6.

Table 3.6: Possible sides that may be viewed in digital representations of microfossils.

ventral:	The ventral (or front) side of a microfossil usually has the most detail and variation out of all the sides.
dorsal:	The dorsal (or back) side of a microfossil is opposite to the ventral side and is usually the smoother or more flat of the two sides.
edge:	This is equivalent to the profile view of a microfossil. In more spherical microfossils it may be hard to tell if the specimen is on its edge. However, for asymmetric microfossils this will be when you can see a difference between the top and bottom silhouette.
oblique:	An oblique view is obtained when a microfossil is tilted. Any of the above view sides could appear at an angle so a foram can be labeled using two words. For example, oblique dorsal implies a tilted microfossil mainly showing its dorsal side.

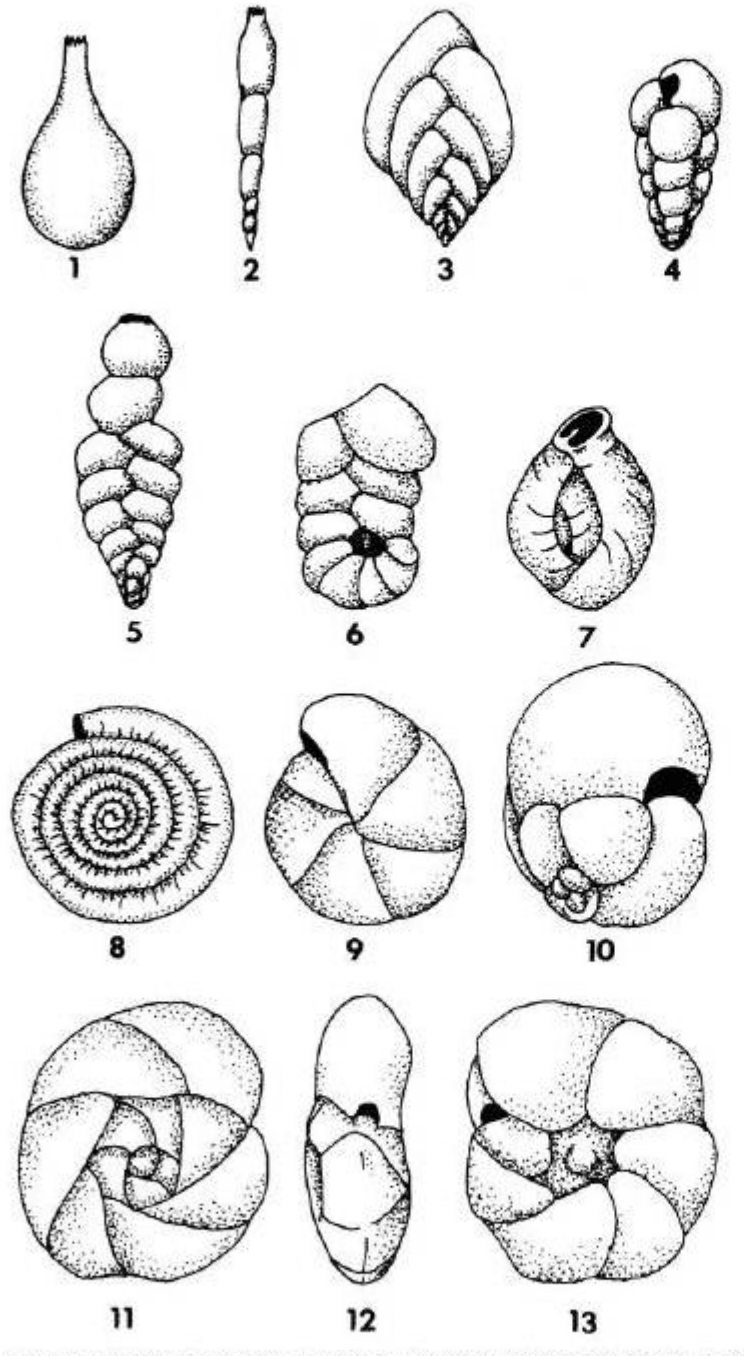


Figure 3.3: Common chamber arrangements seen in forams: (1) single chambered; (2) uniserial; (3) biserial; (4) triserial; (5) triserial to biserial to uniserial; (6) planispiral to biserial; (7) milioline; (8) planispiral evolute; (9) planispiral involute; (10) streptospiral; (11) trochospiral, dorsal view; (12) trochospiral, edge view; and (13) trochospiral, ventral view. Image and names taken from Sen Gupta [80].

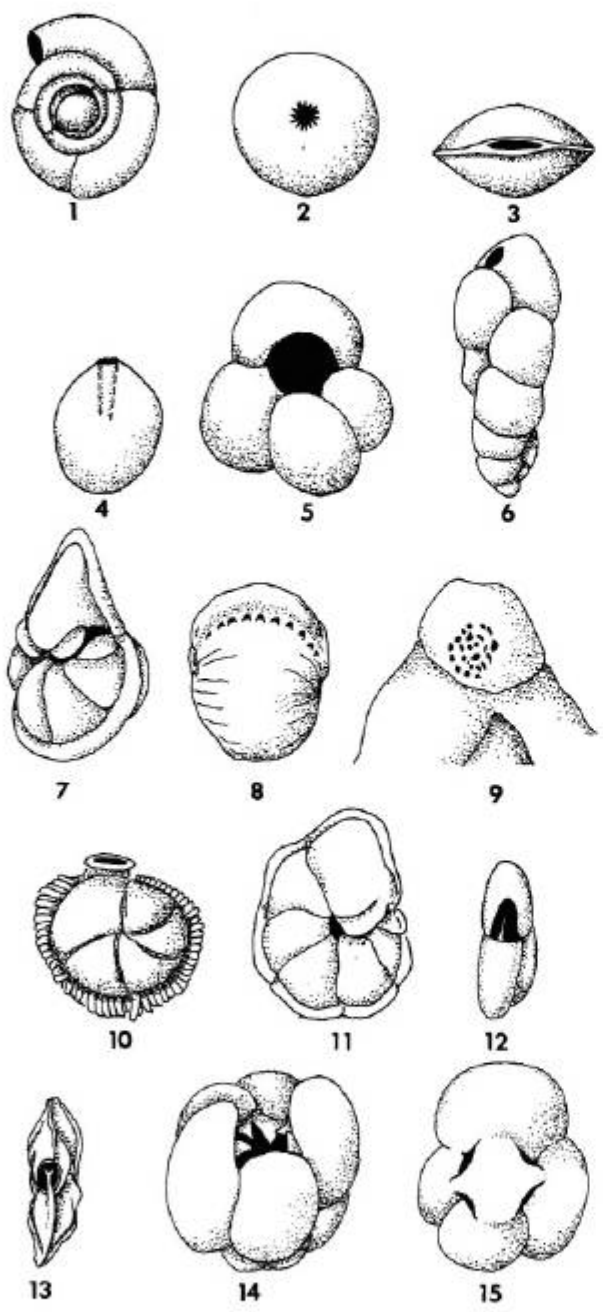


Figure 3.4: Common aperture types found in forams: (1) open end of tube; (2) terminal radiate; (3) terminal slit; (4) terminal with entosolenian tube; (5) umbilical; (6) loop-shaped; (7) interiomarginal; (8) interiomarginal multiple; (9) areal, cribrate; (10) with phialine lip; (11) with simple apertural lip; (12) with simple tooth; (13) with bifid tooth; (14) with umbilical teeth; and (15) with umbilical bulla. Image and names taken from Sen Gupta [80].

3.4 System

The System section is used to give interested users a view of the Microfossil Quest. The details provided are given to encourage others to develop similar systems. A system flowchart shows a high level view of the system modules and how they interact with each other. The system overview page on the website has a simplified version of Figure 3.5, with a detailed flowchart of individual modules shown in linked pages. For this thesis, a single detailed flowchart is given. All references to the individual flowcharts refer to the detailed flowchart. The text seen when users navigate to more detailed information is given in Sections 3.4.1 to 3.4.5.

The Microfossil Quest system is made up of a front-end website and back-end processes. The front-end is this website and the back-end contains algorithms to assist volunteer identification. This section gives a high-level view of the Microfossil Quest system.

Figure 3.5 shows the flow of data in the Microfossil Quest and how the five core modules interact. For better descriptions of each module, click on it in the figure.

3.4.1 Users

The Users module is used in the Microfossil Quest to describe users viewing the identification results obtained from the system. A detailed figure depicting the client and public submodules is shown in Figure 3.5.

The client submodule is used to describe industry and research representatives with the ability to view and submit specimens to the dataset. The final goal is to have a self-sustaining system that is able to generate funds to cover operational and research costs. The commercial prototype will allow industry and researchers to enter specimens into the database for a small fee. These clients can view current identification results for submitted specimens, and choose how they wish to obtain identifications: human intelligence, computer intelligence, or a mixture of both.

The public submodule is used to represent members of the general public interested in viewing specimen results. The target audience for website design are students or graduates from post-secondary institutions. The information available on the website is used to educate and to allow the public to view microfossils that are normally unaccessible for the average person.

Clients and the general public view final results in the specimen dataset. The User module is used to represent these users who see the final results.

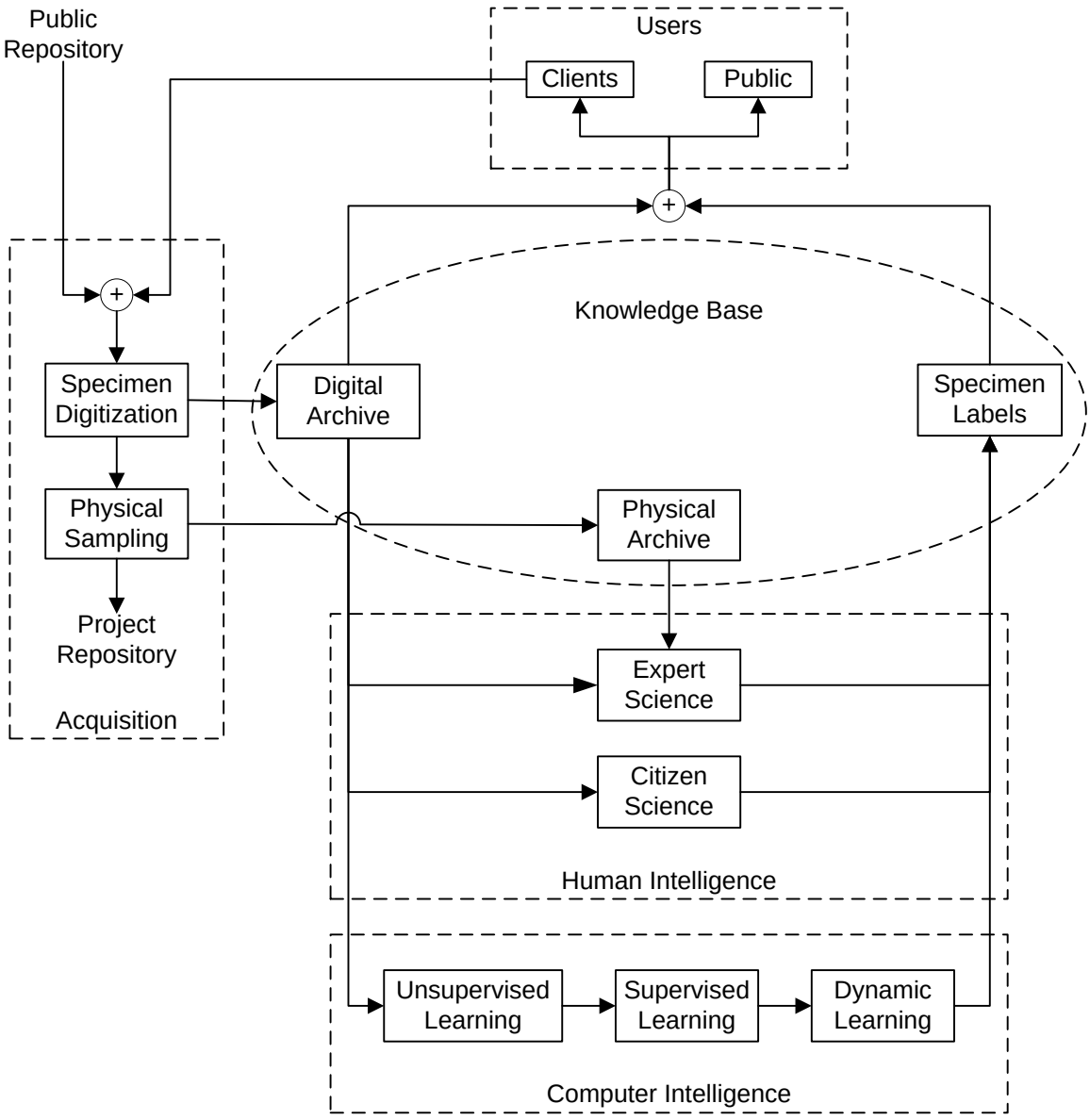


Figure 3.5: High-level view of the Microfossil Quest. How data flows between and within the five core modules is depicted.

Figure 3.6: Vials obtained from a drilling program repository (Figure 3.9). Each vial contains ocean sediment having thousands of microfossils. The microfossils entered into the system are from similar samples to those shown here. Each vial is labeled with information describing the site and depth where the sample was taken.

3.4.2 Acquisition

The Acquisition module is used to take physical specimens and digitize them. Figure 3.5 gives a detailed flowchart illustrating the image acquisition steps: obtaining, preparing, capturing, and storing.

The physical specimens obtained for digitization can come from two separate sources: a repository or clients. Specimens from core repositories are generally ordered or requested by the system developers. These specimens arrive in labeled vials holding samples of sediment, as seen in Figure 3.6, with microfossils ready to be imaged. Microfossils may also be submitted from clients. Clients are able to send microfossils to the developers to be entered into the system. All specimens to be used in the dataset need to be digitized. Digitization is done through the use of several pieces of equipment.

In the specimen digitization submodule, several steps must be followed in order to obtain digital representations. First, a portion of the sediment is sieved into a chosen size fragment. The specimens are placed on a black slide that is then placed onto a stage under a Zeiss Stemi 2000-C microscope (Figure 3.7(a)).

Image capture is conducted using integrated hardware. The stage is made from Micos USA parts with control over x-y-phi orientation. An internal PCI card with motor drivers is used to move the stage horizontally, vertically, and rotationally (Figure 3.7(b)). A light source is used to make all specimens on the stage more visible (Figure 3.7(c)). We control the type of light using a Zeiss KL 1500 LCD fiber-optic light source at a 3050 K colour temperature. Images of each specimen are captured with a PixeLink PL-A622 CMOS microscopy camera (Figure 3.7(d)) mounted on top

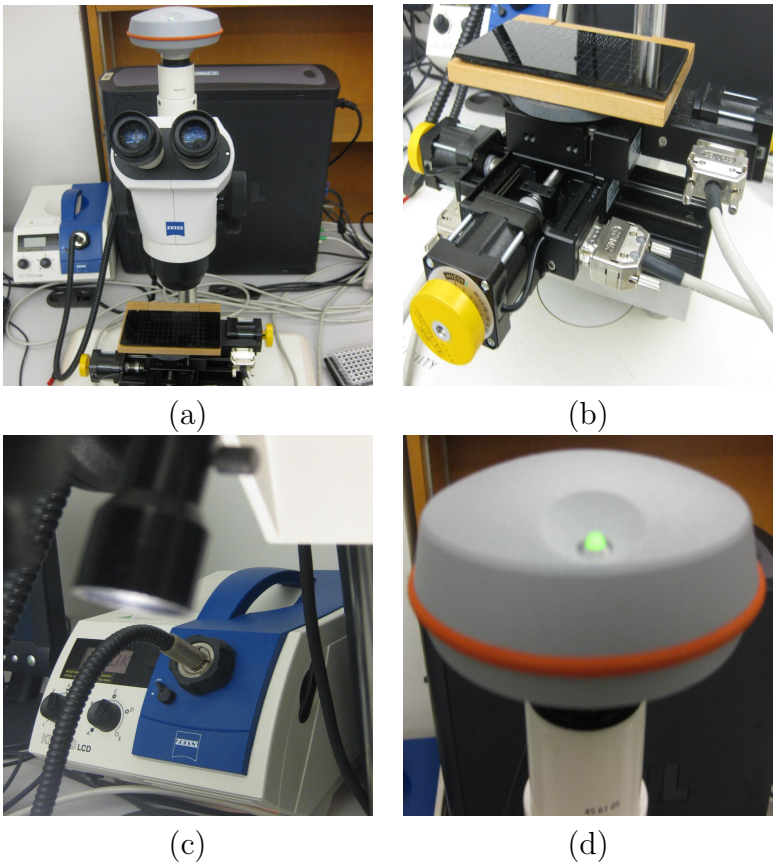


Figure 3.7: Hardware to acquire digital representations: (a) microscope; (b) stage; (c) light source; and (d) camera.

of the microscope. It takes a picture of what a user would see if they looked into one eyepiece in the microscope. Once specimens are digitized, they are automatically loaded into the database along with known information, such as depth range.

The last step in specimen acquisition is to physically archive a subset of the specimens that have been digitized. These are randomly chosen. All remaining specimens are kept in vials in our dataset repository. We do not archive all specimens because physical archival is a time consuming process that is currently done manually. For archiving, a specimen is located on the stage, transferred to a slide and glued thereon. Slides are labeled and labels are entered into the database so that cross-referencing is possible.

Physical specimens are digitized in four steps: obtaining, preparing, capturing, and storing. The Acquisition module accounts for the necessary steps to enter a physical specimen into the system, and to ensure the specimens can be used later for research and validation.

3.4.3 Human Intelligence

The Human Intelligence module of the system is where volunteers view digital representations of specimens from the dataset, and provide identifications. A detailed figure depicting the submodules is shown in Figure 3.5. Volunteers are divided into two categories: trained and untrained volunteers.

Trained volunteers are specialists or experts that are invited to participate in microfossil identification. A specialist has some training in identifying microfossils, while experts are volunteers with formal training to do identifications, usually with a degree in this area. This is considered expert science because knowledgeable users do not require any additional training to provide identifications.

Untrained volunteers are considered novices and citizens. They do not necessarily have any training in scientific research or in micropaleontology. Novices are members that have recently joined the project. Citizen scientists are novice members that have gained a sufficient amount of experience, with relatively good identifications that agree with our experts and specialists. Novice members are upgraded to a citizen level when they tend to generate reliable identifications. In this fashion, the system is an example of citizen cyberscience. Untrained volunteers will need to be trained and educated to perform useful identifications.

Identifications provided by trained and untrained volunteers represent the outsourcing of some computation to humans, which is what makes the Microfossil Quest a human-based computation system. The front-end module is how the system solicits human intelligence, which is leveraged by the computer intelligence back-end of the system.

3.4.4 Computer Intelligence

The Computer Intelligence module in the Microfossil Quest system is meant to serve users but also assists volunteers. This module incorporates algorithms to increase identification performance. As with other crowdsourcing projects, volunteers provide identifications sporadically. We developed back-end processes to ensure a single identification has the greatest impact on the dataset. A detailed figure depicting the submodules is shown in Figure 3.5. System thoroughness is increased through unsupervised learning, reliability is increased through supervised learning, while predictability and throughput are increased using dynamic learning.

The first algorithm is considered an unsupervised approach that groups specimens in clusters using Agglomerative Hierarchical Clustering. By clustering specimens that appear similar, we are able to relate the specimens that are most likely to be from

the same taxonomy. By generating these clusters, we can identify more specimens in the dataset and increase identification thoroughness.

Cluster results use a supervised learning approach to propagate direct identifications (human identifications) using a second algorithm. In this fashion, we are able to take one identification and propagate it through the dataset to generate multiple indirect identifications. Each indirect identification is associated with a confidence rating, which indicates the reliability of the inferred result. The algorithm used was designed to reliably propagate identifications without changing any specimen identifications obtained through Human Intelligence.

Once indirect identifications are known, the third algorithm for dynamic learning was developed to sequence all specimens into a priority list. The priority list encourages volunteers to identify specimens that would improve performance and enable the dataset to be identified quickly through identification propagation. The specimens with the greatest impact on the database are first, and the specimens with the least impact on the database are last. The most important unidentified specimen is considered the specimen that, once identified, would generate the greatest total confidence in the dataset. If all specimens are identified by volunteers, the most important specimen is the one that is most likely to be incorrectly identified, the one we want to double check. Having a priority enables dataset behaviour to be more predictable and the total confidence, or reliability, of the dataset to increase quickly.

Unsupervised learning, supervised learning, and dynamic learning make up the Computer Intelligence currently included in the Microfossil Quest. This Computer Intelligence module represents the computation part of human-based computation and our attempt to improve the reliability and throughput required to fully identify a dataset.

3.4.5 Knowledge Base

The Knowledge Base module in the system contains all the information about each specimen. A detailed figure depicting the submodules is shown in Figure 3.5. The Knowledge Base is separated into three submodules: digital archive, specimen labels, and physical archive.

The digital archive stores all the digital representations of a specimen. In the case of the current Microfossil Quest, this includes non-anaglyph and anaglyph images along with any known specimen information that remains constant for each specimen. The digital representations and known information are received from specimen digitization in the Acquisition module. From here, the representations may be used

to obtain identifications in the Human Intelligence module, or be used in back-end processes to improve identification impact in the Computer Intelligence module.

The specimen labels submodule incorporates all the provided identifications that may change over time, such as taxonomy and feature identifications. This includes the identifications obtained from Human Intelligence and Computer Intelligence. As more volunteers provide identifications for a specimen, the final consolidated identification could change. Once a consolidated identification is determined, it is combined with the known specimen information and viewed in the Users module.

Archived physical specimens obtained from the Acquisition module are also contained in the Knowledge Base. The physical specimens are obtained in order to verify the complete system is working accurately. These archived specimens can be mailed to experts for identification using traditional Particle-Based Identification (PBI) in order to validate system results.

The Knowledge Base is the heart of the Microfossil Quest as it contains the digital archive, specimen labels, and physical archive of specimens in the system. This is the location where all specimens and their associated identifications are stored, altered, and viewed. All other modules in the system interact with the Knowledge Base, making it an essential part of the Microfossil Quest, storing the input dataset and output results.

3.5 Background

To teach citizens about the general history of microfossil research and crowdsourcing, a background section is included. This section is used to briefly describe the different areas involved in the Microfossil Quest. If users are interested in a particular topic and wish to learn more, they are able to click on links to more detailed information given in Sections 3.5.1 to 3.5.3.

The background information that is important to understand the Microfossil Quest is separated into two sections. Microfossil background is important because it is the subject being studied. Crowdsourcing is important to this project because it is the method being used to study microfossils. A summary of both fields is given in this overview, while more details can be found by clicking on the microfossil or crowdsourcing headings.

Microfossils

Microfossils are fossils from any small organism that cannot be seen clearly without a microscope. There are many different types of microfossils all over the world,

but they are most abundant in large bodies of water such as oceans and seas. As microorganisms evolve and change, these evolutionary changes in the organisms are reflected in alterations to their shapes. When microorganisms die, their remains float to the bottom of the water basin and collect as sediment. Over the years, billions of microfossils are preserved in the cold temperatures on ocean and sea floors, forming layers showing this evolution over time. By collecting and studying microfossils we can obtain information about the prehistoric environment on Earth and the present geological factors of a region.

Crowdsourcing

Requesting help from a wide range of people has been conducted over the years and is considered crowdsourcing. An increasing number of projects ask human volunteers to perform tasks. There are many commonly used terms to describe these kinds of projects. Crowdsourcing is a general term covering any type of project using volunteers. Other common terms include citizen science, citizen cyberscience, and human-based computation. The Microfossil Quest project is primarily a human-based computation project and secondarily a citizen cyberscience project.

3.5.1 Microfossils

There are many types of microfossils found around the world in different locations. A good place to find fossils is in the oceans and seas. In the oceans, microorganisms live anywhere from the surface to the ocean floor and when these organisms die their remains drift down to the ocean floor where they collect as sediment. All over the world these microfossils have been collecting on the ocean floor for several hundred million years (Figure 3.8), even before the time of dinosaurs (the Jurassic period).

In order to understand how we came to study microfossils, we delve into the history behind oceanography, a broad field that includes marine micropaleontology. From understanding how oceanography began, we can see how it has influenced worldwide ocean drilling programs, from which we are able to obtain the foram microfossils used in the Microfossil Quest.

Oceanography

Oceanography is the study of the oceans on the Earth. This includes all topics concerning the oceans, such as ocean currents, ocean waves, and marine organisms. Before 1872, there were no purely scientific expeditions for oceanography research. The *HMS Challenger* set out to study the ocean and the ocean floor around the world,

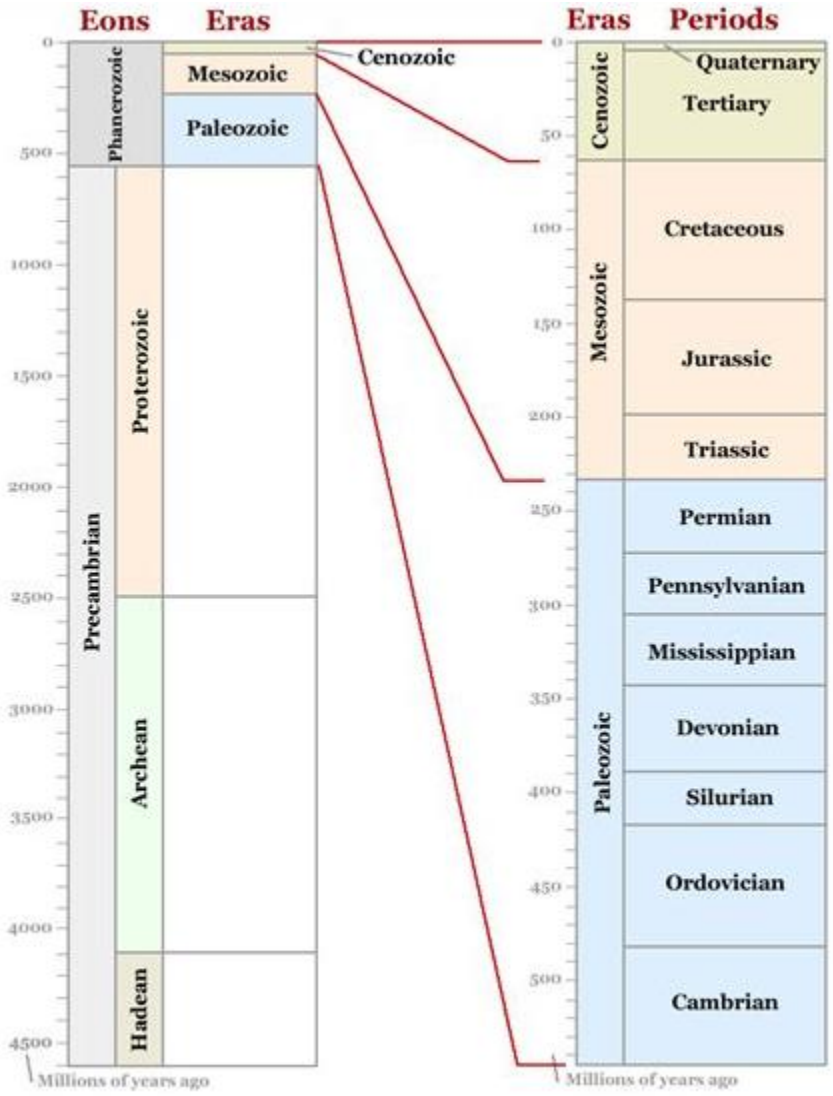


Figure 3.8: The geological time scale used by paleontologists, taken from `geology.com` [81]. The earliest foram fossils have been traced back to the Cambrian period [16–18].

a scientific voyage which had never been done before. Charles Darwin's famous voyage on the *HMS Beagle* (1831–1836) and Thomas Huxley's on the *HMS Rattlesnake* (1846–1850) were not purely scientific missions. These two voyages were primarily for naval purposes. The study of oceanography, as we know it today, began with the HMS Challenger, which changed "the face of science forever" [82].

Challenger was one of the smaller British naval ships—she displaced 2 300 tons, was 200 feet long, and had 15 of her 17 guns removed to make room for labs. In December 1872, Challenger left Portsmouth, England, under the direction of Captain George S. Nares, with 23 naval officers, 240 crewmen, and 6 scientists who were led by Wyville Thomson. The primary goals of the expedition were, as stated by Corfield [82]:

- "To investigate the physical conditions of the deep sea in the great ocean basins in regard to depth, temperature, circulation, specific gravity, and penetration of light;"
- "To determine the chemical composition of seawater at various depths from the surface to the bottom, the organic matter in solution and the particles in suspension;"
- "To ascertain the physical and chemical character of deep sea deposits and the sources of these deposits;"
- "To investigate the distribution of organic life at different depths and on the deep sea floor."

Many things were discovered during the Challenger expedition, including what is now known as manganese nodules, the presence of life far below the ocean surface, and that much of the ocean floor sediment is composed of microfossil shells drifting down from various levels of the ocean above. Challenger was the first expedition whose purpose was for scientific research alone and, in May 1876, Challenger finished the expedition. It has pioneered the way for other similar expeditions and studies, such as modern ocean drilling programs, of which the first used a ship named the *GLOMAR Challenger* [82].

Drilling Programs

There have been three international ocean drilling programs to collect sediment cores. The programs are the Deep Sea Drilling Program (DSDP) of 1986–1983 [12], the Ocean Drilling Program (ODP) of 1985–2003, and the Integrated Ocean Drilling

Figure 3.9: A collection of ocean core samples from the IODP Bremen Core Repository, taken from the IODP website [13]. Researchers are able to order samples from these ocean cores for study.

Program (IODP), which started in 2003 [13]. All cores that were collected are stored in various repositories waiting to be used. An image of the IODP Bremen Core Repository is shown in Figure 3.9.

Samples from these cores can be extracted into vials and sent to researchers around the world. These vials contain vast quantities of microfossils that can be used to study biostratigraphy and prehistoric environmental conditions.

Forams

One microfossil order that is commonly studied by micropaleontologists is called foraminifera. Foraminiferida are small unicellular organisms that make shells [14]. Live foraminiferida evolve rapidly with different species located at different ocean depths. Fossilized shells are called foraminiferal tests, or forams.

Forams are examined to aid with biostratigraphy, paleoclimatology, and paleoceanography. Biostratigraphy uses the taxonomic analysis of fossils in a rock layer to determine the relative rock age. In paleoclimatology and paleoceanography, geochemical analysis is used to determine the elements and isotopes in the forams [14, 21, 22].

This information is used to study prehistoric atmospheric carbon dioxide levels, global carbon cycle, and ocean temperatures.

The importance of, widespread use of, and ease in obtaining forams is the reason why the Microfossil Quest system focuses on these microfossils. Obtaining a large dataset of identified forams would be beneficial to researchers studying prehistoric environmental conditions and for industrial applications including biostratigraphy.

3.5.2 Crowdsourcing

Developing projects incorporating human interaction is a new approach gaining popularity. Many crowdsourcing projects are linked to various names, including citizen science, citizen cyberscience, and human-based computation. Citizen science is described as the use of volunteers to help perform scientific research while educating volunteers on a scientific area or the scientific process. These tasks may range from collecting data (e.g., taking pictures of things) to performing analysis (e.g., describing what is seen in the pictures) [38]. Volunteers may be asked to perform tasks for many reasons. For example, a research group may not have the time or staff to do these tasks themselves. Citizen cyberscience is considered a subset of citizen science that involves computers and/or the Internet as a major component. Human-based computation projects request humans to perform certain actions, with the results further processed to complete the main goal of the system. Humans assist computers with a task. The Microfossil Quest is considered a human-based computation project that utilizes citizen cyberscience in order to obtain microfossil identifications.

The Christmas Bird Count held by the American National Audubon Society, Games With A Purpose (GWAP), reCAPTCHA, and Galaxy Zoo are popular examples of crowdsourcing projects. The Christmas Bird Count is a citizen science example, while GWAP and reCAPTCHA are good examples of human-based computation projects. Galaxy Zoo is an example of a combination of citizen cyberscience and human-based computation. It shows how one project can fall under multiple crowdsourcing categories.

American National Audubon Society

The American National Audubon Society (ANAS) [41, 42] provides an example of citizen science that is not citizen cyberscience. ANAS volunteers count birds in order to determine where birds are migrating and in approximately what numbers. This project, called the Christmas Bird Count, actively counts birds annually for two weeks around the end of December and beginning of January. The Christmas Bird Count

has been running since July 1900 and had a total of 59 918 observers who counted 57 704 260 birds from December 14, 2007 to January 5, 2008 [42].

Due to the success of the Christmas Bird Count, the Backyard Bird Count was started and has been running for thirteen years. The Backyard Bird Count is held for four days in February and has met with success, receiving 92 206 checklists and counting 11 471 322 birds in 2011 [44].

GWAP

Games with a purpose is a human-based computation initiative asking users to perform tasks in a game, which provides information that can be used for further purposes [48]. The ESP game, used to provide labels for online images [47], is the first of these types of games. Another example is Verbosity, which was developed to obtain common-sense facts [59]. The Verbosity clue structure was created to encourage users to provide facts that could be easily incorporated into the Open Mind Common Sense (OMCS) project [60]. By comparing Verbosity data to ConceptNet data—where sentences are automatically generated using OMCS results—it was determined that "the speed of knowledge acquisition... is much faster than the standard volunteer-based interface" [60]. By incorporating processing to "weed out data that will not be useful for the target system" results could be improved, while maintaining the gaming aspect to motivate users [60].

reCAPTCHA

reCAPTCHA is a human-based computation project designed to assist with the transcription from physical books and text to digital books [57]. reCAPTCHAs are based on Completely-Automated Public Turing Test to tell Computers and Humans Apart (CAPTCHAs), which are distorted text that is difficult for computers to translate but easy for humans. By 2008, reCAPTCHAs were used by more than 40 000 websites around the world to validate if a user is a human or computer [57]. If you were ever asked to look at an image and type the visible text seen in the image, you are providing CAPTCHA or reCAPTCHA information. Typically, Completely-Automated Public Turing Test to tell Computers and Humans Apart (CAPTCHA)s have one block of text for human validation, while reCAPTCHAs use two text blocks, one needing translation, and a second word for CAPTCHA purposes. Because of the difficulty for computers to process and recognize these texts, reCAPTCHA is a reliable security measure to prevent "large scale abuse of online services" [57]. But reCAPTCHA is also able to digitize books, which is the main goal of the project.

Galaxy Zoo

Galaxy Zoo is a successful citizen cyberscience and human-based computation project [55]. In this project, volunteers originally helped label galaxies photographed by the Sloan Digital Sky Survey [54,55]. In their first published paper, the Galaxy Zoo team stated they had 85 276 users giving 893 212 galaxy identifications. From all these user identifications, 39 distinct identifications for each galaxy were obtained after screening and filtering of user identifications [54]. They have since expanded to five citizen cyberscience projects using images taken from NASA's Lunar Reconnaissance Orbiter and the Hubble telescope. The success of this project was a major inspiration for the Microfossil Quest project.

3.5.3 References

References are included in the website to allow us to cite our sources of information. Including the references also provides interested volunteers with locations to research more information into a topic and verify the facts mentioned throughout the site.

All the references found throughout the Microfossil Quest website can be found on this page.

A list of all the references used throughout the website follows. As they are a subset of the references given in this thesis, they are not repeated here. When any source is referenced in the website, there is a link redirecting the browser to this page.

3.6 Conclusion

Human interaction is an important component of the Microfossil Quest because identifications are obtained from volunteers through a website. By allowing citizens to volunteer and provide information, we are able to leave the most complex and difficult tasks of microfossil identification to humans. Asking citizens for help enables us to obtain results faster than if we only approach a select number of specialists or experts with limited time availability. When developing the website, a new approach to specimen identification was used, better digital representations of specimens was integrated, and a large amount of educational material, covering various aspects of the system, was incorporated.

Unlike most citizen cyberscience approaches, the method volunteers use to identify specimens was altered. In many other approaches, images and objects to be identified by volunteers are chosen for them. In our case, we use a newer method, partially implemented by Herbaria@home, where volunteers are able either to identify the

default specimens or to search the database for preferred specimens to identify. In our system, we allow much more control over the specimens volunteers can identify by integrating the specimen search with our identification. By giving a significant amount of control to volunteers, citizens can focus on providing taxonomic and/or feature identifications with which they are comfortable.

To make the website more interesting, and to provide better digital representations of specimens, the VRLM applet was created. The VRLM applet allows users to view anaglyph or non-anaglyph images and to vary illumination conditions, which results in more appealing and informative digital representations. A similar applet, modeling illumination and depth information for a physical object, has not been used before in other citizen cyberscience projects.

Incorporated into the Microfossil Quest website is a large amount of educational material. Some of this material is of the kind normally seen in citizen cyberscience projects, but few projects include all the kinds of information available on the Quest website. Volunteers are able to learn about microfossils and forams through the tutorial, which goes into detail about defining taxonomy features. Volunteers and users also get a high-level view of the software design through the system section, where the Microfossil Quest system is described. A summary of the background to marine microfossil research and crowdsourcing is also provided to volunteers. Along with the basic identification and citizen motivation sections, there is a variety of detail and information included in the human-interaction component of the Microfossil Quest system.

Many new features were incorporated during the development of the Microfossil Quest website: these include the combination of search and identification; the development of the VRLM applet; and the inclusion of educational material describing training, background information, and the system. Preliminary website flow, interaction structure, and the database itself have been developed. In addition to a complete draft of the website text, the overall functionality and base features have been created. With crowdsourcing supported by the front-end website, the remaining component for the Microfossil Quest system is the back-end processing or computation algorithms, which are described in Chapter 4.

Chapter 4

Computation Algorithms

The end goal of the Microfossil Quest system is to develop a fully-automated microfossil identification system. In order to achieve this goal, we have been using the evolutionary prototyping design model. For our latest prototype, we have gone from a computer-aided approach to a human-based computation system. The human-interaction front-end of the system is designed to collect identifications. The computation algorithms in the back-end leverages the Microfossil Quest system to meet our application requirements. In many crowdsourcing projects, computational processing is included according to the needs and goals of the system. Data may be obtained in crowdsourcing projects from only humans, from computer-assisted humans, or from humans and from computers.

Many crowdsourcing projects obtain data strictly from humans. Developers focus on human sources for identifications because tasks are too difficult for computers and too time consuming for researchers to complete. Examples of this are Galaxy Zoo, Herbaria@home, and Stardust@home. All three of these projects depend on obtaining identifications solely from humans. In the case of Galaxy Zoo, the project relies on having a large amount of independent identifications to ensure reliability [72]. Herbaria@home obtains initial identifications from volunteers. Validation of the results is also done by volunteers, including users of the website who are able to change the results any time they notice an error [45]. Startdust@home is slightly different because all aerogel tiles obtain initial identifications from volunteers and project developers verify identifications results [51]. In all these projects, the original data and its verification is done by humans. In the case of Galaxy Zoo and Stardust@home, identifications are difficult to automate, while in Herbaria@home not only image processing but also background knowledge must be associated with certain image patterns, which currently cannot be done by computers. For these reasons, leaving identifications to humans is ideal in these projects.

In other crowdsourcing projects, identifications are obtained from computer-assisted humans. One example of this can be seen in the ESP Game. In the ESP game, two players type words that come to mind when looking at an image. If they type the same word, the label is applied to the image [47]. Once X number of player pairs agree upon the same label, it is then considered a taboo word that future players are unable to use. This incorporation of taboo words leads volunteers to determine a variety of labels for a single image. A single image has been fully labeled when the image has acquired a list of taboo words such that new players repeatedly request to pass on the image because they are unable to think of, or agree on, new labels. The computation back-end of the ESP Game decides both when labels are considered taboo and when an image has been fully labeled, but the labels for each image are determined by volunteers. Due to the difficulty to generate labels automatically, and the variety of possible labels for a single image, the incorporation of simple computation algorithms ensures images receive as many labels as appropriate.

The last, more complex, approach is to obtain identifications from both humans and computers. reCAPTCHA is a good example of a crowdsourcing project where identifications are obtained from two sources [57]. reCAPTCHA translates images of physical books into digital text. To do this, words in each image are segmented and identified by two Optical Character Recognition (OCR) programs. If these two programs do not agree on a segmented word, it is considered an 'unknown word' for humans to identify. Valid human identifications for unknown words are considered one vote for the volunteered translation, while the translations generated from the OCR programs receive half a vote. Translations receiving a total vote of 2.5 or more become the accepted result. In reCAPTCHA, identifications are obtained from both human and computer sources. The OCR algorithms attempt to identify all the words in the book to be digitized, with only a small subset of words being sent to volunteers. This has the advantage of quickly processing a large amount of data, while ensuring reliability through the incorporation of volunteer identification when the OCR results are in doubt.

A wide variety of identification methods can be seen in crowdsourcing projects from human identification, to computer-assisted human identification, to joint human and computer identification. In comparison, computation algorithms created for the Microfossil Quest are complex examples of human-assisted computer identification. In this method, the computer determines identifications reliably and dynamically for unknown specimens based on taxons applied to some specimens by volunteers. This new approach is designed to address the requirement for a high quantity of specimens to receive high quality identifications quickly. The algorithms developed to meet the

Microfossil Quest requirements are described in Section 4.1. Once the algorithms were developed and implemented, tests on performance were conducted with results presented in Section 4.2. Implications of the algorithms are discussed in Section 4.3, ending with the conclusion in Section 4.4.

4.1 Method

The computation back-end of the Microfossil Quest was designed to leverage the system to identify a high quantity of specimens, while obtaining high quality identifications, quickly. From this requirement, we associate quantity with thoroughness, quality with reliability, and quickly with throughput. To satisfy the requirement, we analyzed what algorithm to use as the base for our design. Focus was placed on clustering algorithms because the Computer-Aided System for Specimen Identification and Examination (CASSIE) prototypes, described in Chapter 2, demonstrated the ability of clustering to compromise between relative effort and reliability. To determine suitability for the Microfossil Quest, ANN, k-means, KNN, and AHC methods were considered.

With Artificial Neural Network (ANN) approaches, a set of weights and training data is used to separate clusters. In this type of approach, training data must be known before-hand with sufficient data to distinguish all possible outcomes. Due to the difficulty obtaining this training data, an ANN clustering approach would not be suitable for this system.

K-means is another clustering method that could be used in the Microfossil Quest. K-means clustering has the benefit of not requiring any known identifications. However, it introduces an additional application requirement for predictability. In order to assess the dataset to ensure reliability is maintained while increasing throughput, there must be some predictability in algorithm behaviour to ensure increased throughput does not compromise reliability. K-means clustering randomly generates patterns for the center of clusters and these central patterns are updated depending on what images are placed in the cluster [83]. Because of this randomness and an inability to assess the reliability, k-means clustering was not chosen.

K-Nearest Neighbour (KNN) is a popular and standard algorithm for clustering. KNN requires the dataset to be separated into a known and unknown subset. Every unknown specimen is identified by taking the majority identification from neighbours [84, 85]. KNN has the benefit of being a simple and standard approach to clustering specimens, but is generally designed for non-hierarchical clustering. In addition to this, clustering is normally conducted when there are fixed known and unknown

subsets. However, crowdsourcing projects are dynamic with the dataset constantly updated. Depending on the sequence of identifications performed by volunteers, the reliability of generated identifications from clustering may be impacted. This makes identifications generated by a KNN approach unpredictable.

This leaves our previous Agglomerative Hierarchical Clustering (AHC) algorithm used in CASSIE 2. The AHC approach is an ideal crowdsourcing approach because it creates a tree, or dendogram, while clusters are being formed. AHC is an unsupervised approach so known identifications are not required to perform clustering. The tree shows the formation of each cluster and how specimens form clusters based on their image similarities. Having a hierarchical representation depicting how clusters are formed allows for the development of algorithms using cluster formation as opposed to just final cluster results. This allows for different approaches to ensure thoroughness, reliability, throughput, and predictability.

After examining ANN, k-means, KNN, and AHC-based clustering approaches, we determined that modifying the AHC approach would yield the best results for our application. In order to maintain the thoroughness, reliability, predictability, and throughput requirements, a new algorithm called Dynamic Hierarchical Identification (DHI) was developed based on AHC. The full DHI algorithm can be subdivided into unsupervised, supervised, and dynamic learning parts. A description of the unsupervised learning part is given in Section 4.1.1. The supervised learning part is described in Section 4.1.2. Lastly, the dynamic learning part is described in Section 4.1.3.

4.1.1 Unsupervised Learning

Digital representations are used to form clusters in the unsupervised learning part of the DHI algorithm. Clustering was incorporated into the Microfossil Quest back-end to increase the thoroughness of dataset identifications. With specimens in clusters, we will be able to propagate identifications ensuring more specimens in the dataset obtain an identification. To use the system, specimen images are captured with some initial preprocessing to prepare the input before the clustering algorithm is run.

Because we have ground truth identifications in the data collected for the CASSIE 1 prototype, we use the same similarity metric and images captured during the testing and validation of CASSIE 1. Specimens were sieved, sprinkled onto a black glass slide, and images were captured. Once captured, images were then processed as specimens can be positioned and oriented randomly, which affects system calculations and performance. To account for this, an invariant transform is used. It starts by segmenting

images to generate a binary silhouette showing the outline of the specimen. Using this silhouette, principal component analysis is conducted and a transform is defined to normalize the image. Normalization is conducted on specimen rotation and position, along with image size. Once the images are ready, visual similarity between all specimen pairs is calculated using correlation coefficients of the processed images. More detailed information describing the specimen acquisition, normalization, and similarity calculation, including the correlation-coefficient equation is given by Ranaweera *et al.* [65]. Now that we have acquired the similarity ratings between all pairs of specimens, we are able to begin clustering.

Cluster formation is conducted using the AHC method. To illustrate how the unsupervised learning algorithm behaves, we give an example of the AHC algorithm and how the trees are formed; pseudocode for this algorithm may be found in Appendix C. Our clusters are formed using the well established AHC method [86,87], as illustrated in Figure 4.1. At the beginning, each node is a cluster with one specimen, as seen in Figure 4.1(a). The pair that is most similar gets combined into a new cluster, as shown in Figure 4.1(b). In this example, specimen 2104 and 2105 have the highest similarity. When the two clusters are combined, the new cluster must consolidate all remaining similarities. In Figure 4.1(a), specimen 1633 has a similarity of 0.3066 to specimen 2104 and a similarity of 0.3122 to specimen 2105. These two similarity ratings must be consolidated. When combining the similarity pairs, the weakest similarity is kept. Using the example with specimen 1633, the weakest similarity of 0.3066 is kept, as seen in Figure 4.1(b). This is repeated for all similarities connecting to the new cluster. Once we reach this new state, we repeat the algorithm, looking for the most similar cluster pair. Eventually, all the clusters will be combined into one big cluster. The code is designed to run until only one cluster is left, enabling us to find the clustering state at any iteration.

To make it easier to view, analyze, and use the information generated from the unsupervised algorithm, we make a tree. In the tree, each node represents a cluster. The tree shown in Figure 4.2 shows how the example seen in Figure 4.1 is visualized. Because the clusters all start with one specimen, the specimen ID is given on the leaves of the tree. As the leaves are combined, the count for the number of specimens in the cluster is shown. The merge levels in the tree indicate the minimum similarity score within the clusters. All clusters that merge closer to 1.0 on the y-axis enjoy more visual similarity of specimens, which would imply a greater chance for the specimens in the clusters to be of the same species.

To understand how the unsupervised part of the DHI algorithm behaves, we gave an example describing how AHC clustering is conducted. This method of cluster-

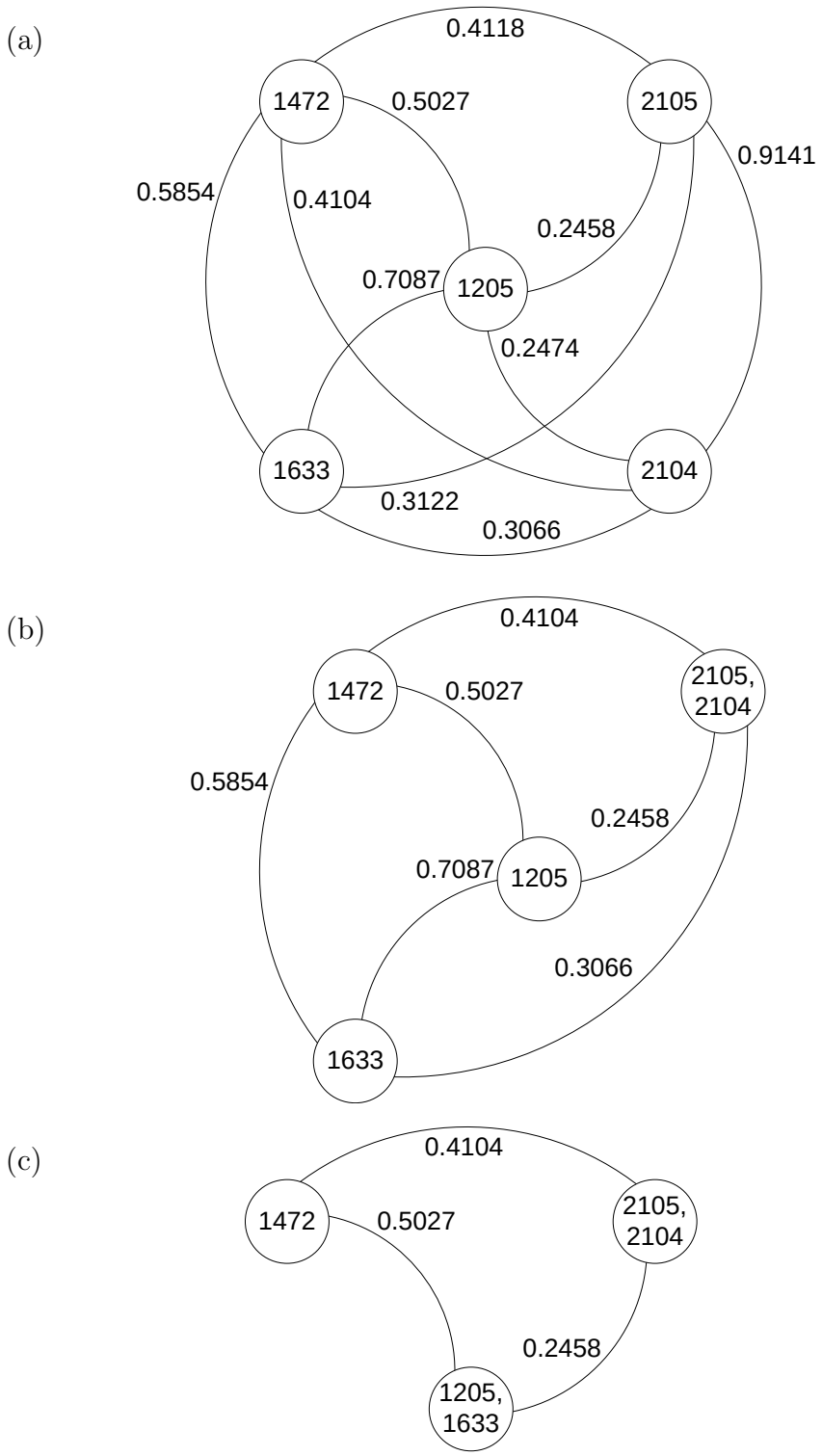


Figure 4.1: AHC nodes indicate clusters and lines indicate similarity scores, which range from zero (uncorrelated) to one (perfectly correlated). (a) Original graph along with similarity pairs. (b) After the first clustering step. (c) After the second clustering step.

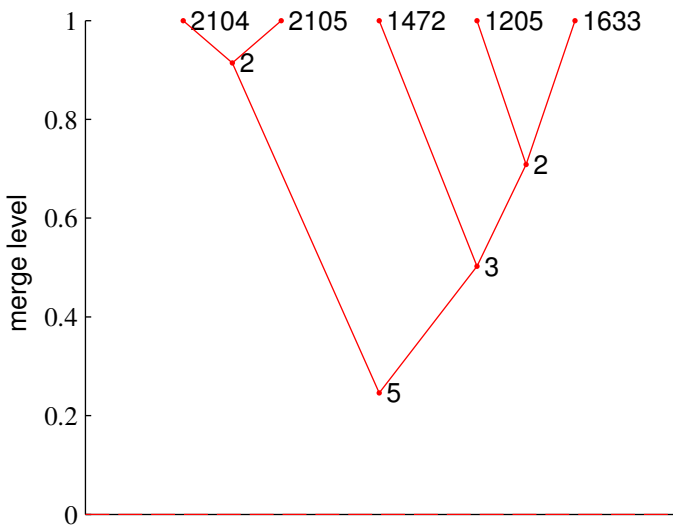


Figure 4.2: A tree representation, which depicts the formation of clusters using AHC until no further clusters can be formed. When each cluster is formed, all specimen pairs therein are similar by at least the merge level.

ing does not depend on obtaining direct identifications. At this stage, clusters are generated through the digital representations that are obtained when specimens are first entered into the database. Now that clusters are known, specimens are related to each other so a single identification can be propagated to many specimens. This relation ensures thoroughness in system performance. The next step is to generate indirect identifications. The Microfossil Quest relies heavily on the generated trees to determine how identifications are propagated in a supervised approach.

4.1.2 Supervised Learning

A supervised algorithm was developed to propagate identifications reliably using the tree from the unsupervised algorithm. Generated identifications for specimens using the supervised algorithm are considered indirect identifications, while identifications provided by volunteers are direct identifications. This propagation is considered supervised because it requires direct identifications before any indirect identifications can be generated. The propagation algorithm, identification confidence, and multiple tree generation is described.

Identifications are propagated according to cluster formation as visualized by trees. Figure 4.3 shows an example of how identifications are propagated. The pseudocode for this algorithm is provided in Appendix C. At each merge level, the specimens in the cluster are examined and unknown specimens are identified, while known specimens are left alone. If any specimen is unknown, the most probable known identification—the majority—for that cluster is used as the indirect identification for unknown specimens. The identification confidence for the indirect identifications is set to the value of the merge level. If there is a tie for the most probable identification, a random choice is made. Looking at Figure 4.3 at merge level 0.9, a cluster of two specimens—with one known identification, *M.vela*, and one unknown identification—results in the unknown specimen receiving an indirect identification of *M.vela* with a confidence of 0.9. At merge level 0.108, when all the specimens are combined into one cluster, one unknown is remaining. This cluster contains three *M.vela* identifications and two *M.subb* identifications so the majority identification is *M.vela*, and this is given to the unknown specimen with a confidence of 0.108. When the propagation is complete, all direct identifications are given a confidence of 1, while any remaining unknown identifications get a confidence of 0.

Confidence levels give an indication of how reliable the indirect identifications are. This also enables users to get an idea of the reliability of identifications in the full dataset by examining the average confidence in the dataset. Confidence

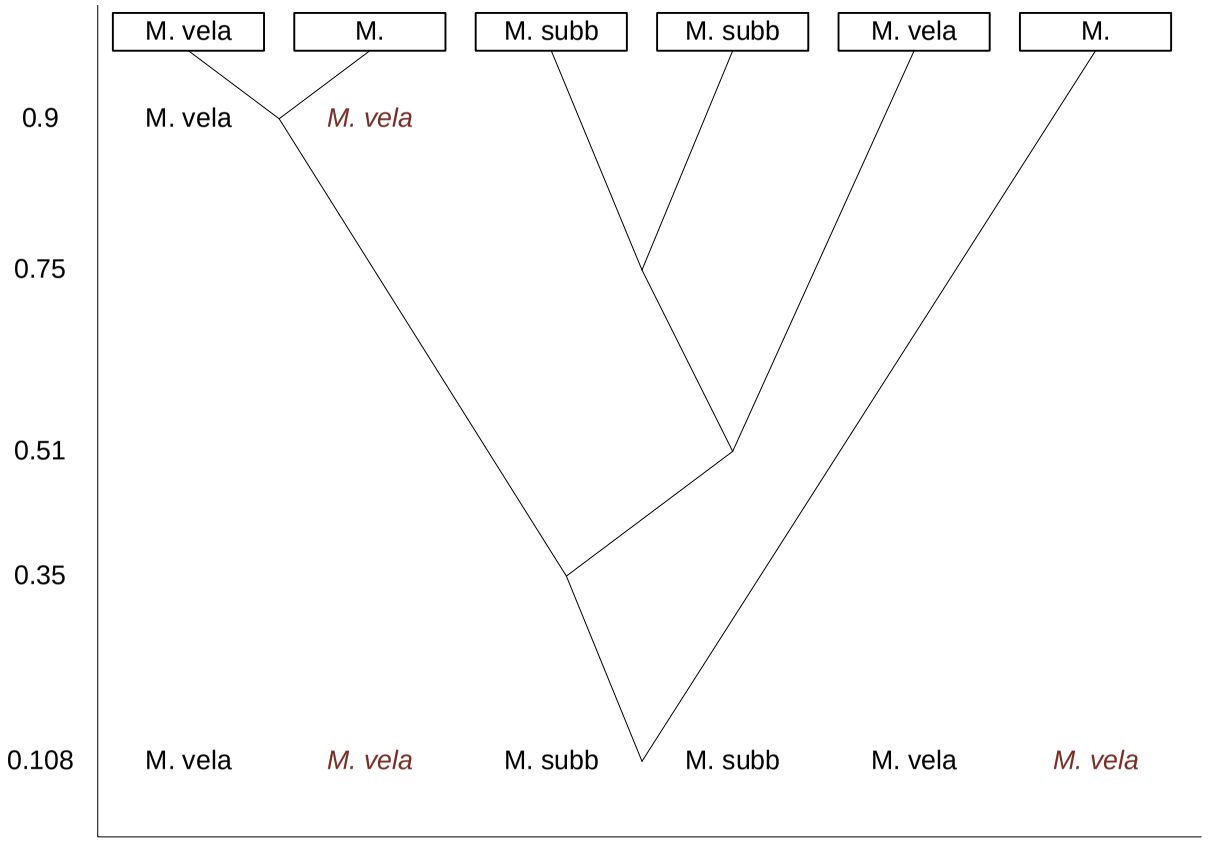


Figure 4.3: This example shows how identifications are propagated at the species level. The left side indicates the merge levels, which reflects the similarity scores used to combine clusters.

levels allow for more control over identifications. If users desire identifications to be propagated as long as a minimum confidence is met, then propagation can be halted at the corresponding merge level. This gives users more control over the clustering and propagation. The trade-off for reliability is throughput and thoroughness. If identification propagation is halted at a high confidence, fewer indirect identifications will be generated. With few indirect identifications more direct identifications will be required to obtain a fully identified dataset, which will take time to obtain.

The hierarchical nature of Linnaean taxonomy means tree generation is done for each taxonomic level. We use multiple trees because taxonomic identification of species should be mutually exclusive between genera, and the specimens included in a tree will affect the merge levels for the clusters in the tree. As an example, Figure 4.4 shows the tree generated at the genus level and Figure 4.3 shows a subset of the specimens for identification at the species level. In Figure 4.3, the species identification tree has been filtered to include only specimens with direct and indirect identifications of genus *Morozovella*. Similarly, Figure 4.4 has itself been filtered to include only specimens of order foram, determined both directly and indirectly. When compared, it can be seen that some merge levels are different. The different merge levels occur in circumstances when clusters are merged and remaining similarity ratings yield different results. Different clusters may be formed potentially yielding different identifications or confidence values. As this may have a high impact on identification propagation, multiple trees are generated to prevent errors. Due to the hierarchical nature of taxonomic identification, tree generation must be completed from highest taxon level, order, to lowest taxon level, species.

A supervised learning algorithm for identification propagation was described before identification confidence was introduced and the need for multiple trees was examined. By using the unsupervised tree, we are able to dynamically propagate identifications at any time and do not require a set amount of known specimens. The goal of this algorithm design was to ensure identifications are propagated as reliably as possible, independently of the amount of identifications that are known. With the ability of this algorithm to generate confidence levels in each identification, the next stage is to ensure direct identifications have the biggest impact on the dataset.

4.1.3 Dynamic Learning

Volunteer participation will vary from month to month, as noted in the Galaxy Zoo project [54], so it is important to ensure all direct identifications have the greatest impact. To leverage human intelligence and increase throughput without reducing

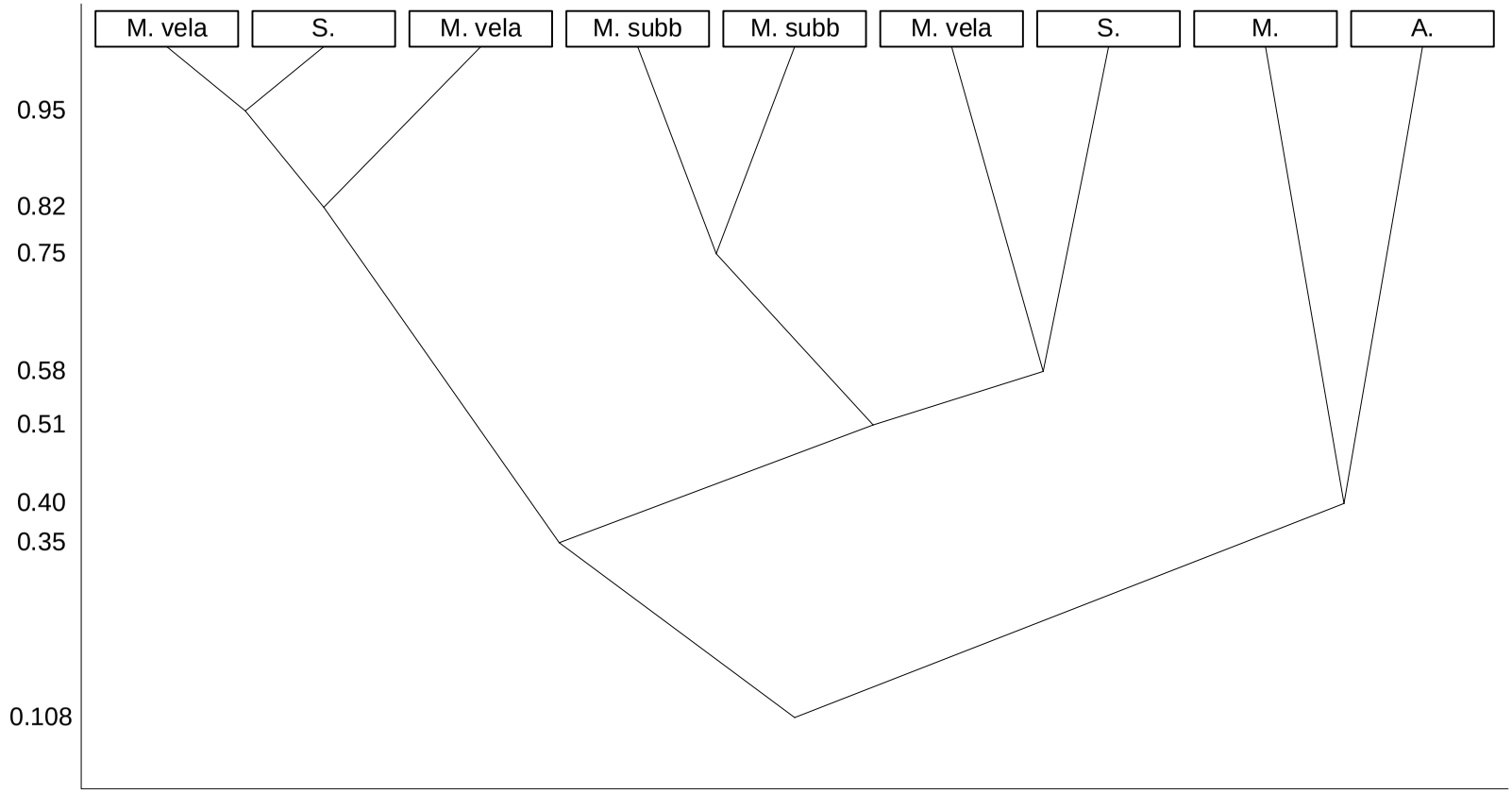


Figure 4.4: This tree was constructed for identification at the genus level. If you compare this tree to the species-level one in Figure 4.3, which is exclusive to one genus, you can see there are some alterations in the merge levels because of the extra specimens.

Table 4.1: Levels for priority generation and the taxonomic hierarchy to be examined at each level.

Level	Order	Genus	Species
1	unknown	unknown	unknown
2	known	unknown	unknown
3	known	known	unknown
4	known	known	known

reliability, a greedy priority algorithm for dynamic learning was developed. This algorithm was used to locate the single specimen that will have the greatest impact on the dataset after identification. A high impact is defined according to the greatest increase in average confidence in the dataset. The focus of the prioritization part of the DHI algorithm is to arrange the specimens in a sequence that would enable the supervised learning algorithm to quickly increase the average confidence level of all specimens in the dataset. We describe the ideal algorithm before describing the practical algorithm behaviour for unknown specimens, known specimens, and multiple trees.

As with the clustering algorithm, priority generation also runs in a hierarchical manner. The levels are shown in Table 4.1. Levels are decided based on the impact on the dataset. Identifying a specimen with no identification whatsoever would have a greater impact than identifying a specimen that has been partially identified. Ideally, the best way to prioritize would be to generate a tree and to locating the unknown specimen that would have the greatest impact on average confidence level in the dataset if it was identified directly. This specimen would have the top priority. The algorithm would then repeat with this specimen now treated as known, generating a new tree each time and adding to the priority list. This would be a very time consuming process so we have devised a way to calculate a priority list that would be similar to this ideal algorithm using a snapshot of the database information and one tree for each unique group at each level.

An example is used to illustrate this algorithm, with pseudocode available in Appendix C. To describe how unknown priorities are generated, we begin by explaining priority generation for the first level in Table 4.1. To generate the new priority for specimens, distances are calculated using the merge levels ($priority = 1.0 - merge\ level$) to indicate how much improvement would be made to total confidence if the specimen was identified. Figure 4.5 shows an example of how priority values are set for specimens. Starting at the leaves, the first merge level is at 0.9. Here, two specimens have the highest priority. The sharpest specimen, as calculated

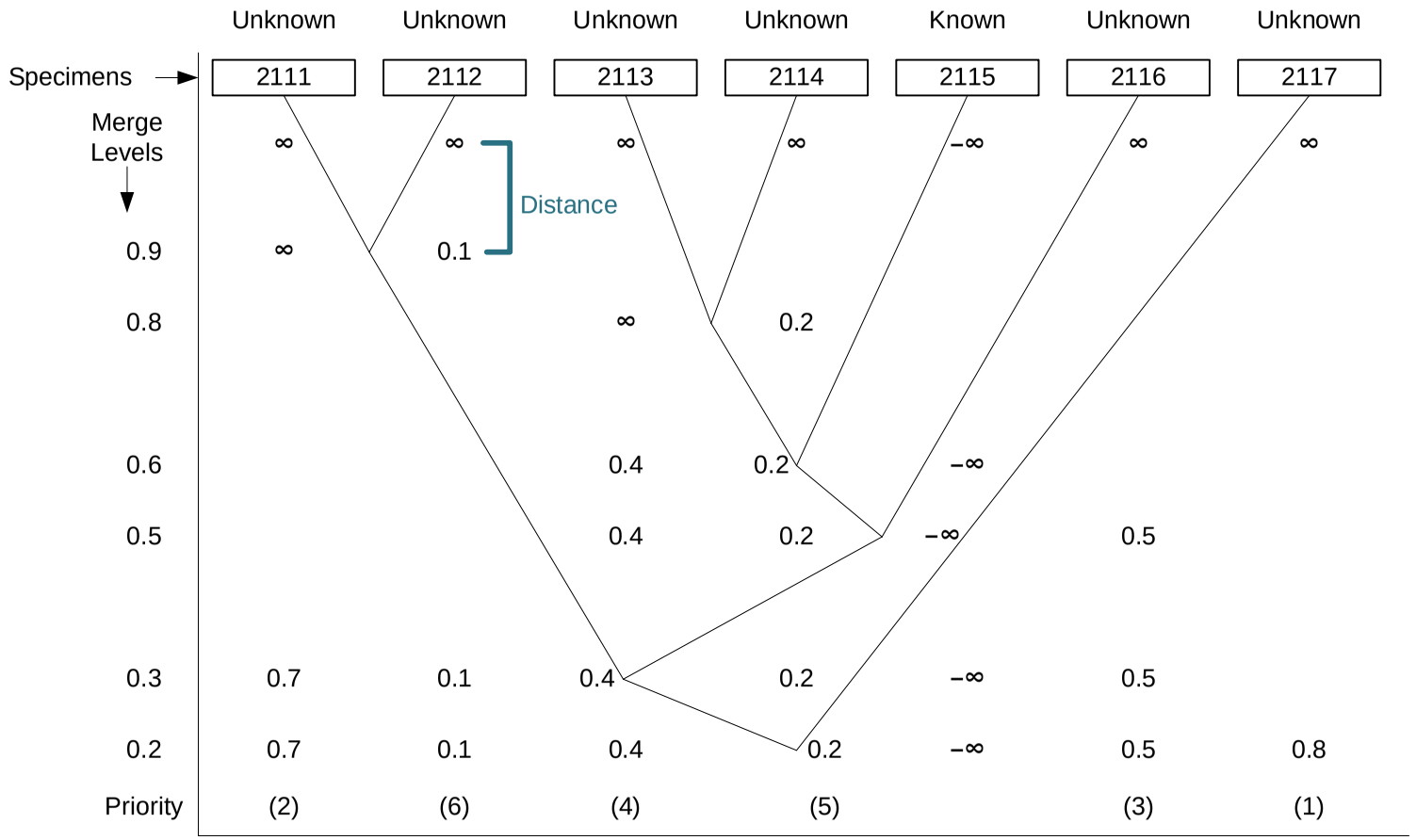


Figure 4.5: This example shows how unknown specimens are prioritized. In this case, fully unknown specimens are used to describe how order priority is generated using distances. At each merge level, all relevant distances are examined and a new priority is calculated if there are any conflicts or if there is a known specimen.

in Appendix B, is left with the highest priority (infinity) while the remaining specimen (2112) is given a new priority of 0.1 ($0.1 = 1.0 - 0.9$). At merge level 0.6, the cluster has a specimen with a known identification so each unknown specimens will be given a finite priority. In this case, specimen 2113 is given a priority of 0.4. Merge level 0.3 is a combination of two clusters. The first cluster has specimens 2111 and 2112, and the second cluster has specimens 2113, 2114, 2115, and 2116. In the first cluster specimen 2111 has top priority, while in the second cluster specimen 2116 has a known identification. Using our algorithm specimen 2111 would be given a new priority. At the end, once all specimens are prioritized they are sorted in decreasing order, and all the known specimens (priority of negative infinity) are removed. The final priority list for this hierarchy would be 2117, 2111, 2116, 2113, 2114, then 2112.

For the fourth level in Table 4.1, where all specimens are fully identified, a slightly modified priority generation approach is used. Instead of soliciting identifications to boost confidence, as no new identifications are needed, we check specimen clusters at low merge levels because they are more likely to be clustered inaccurately, and more likely to have a variety of taxonomies in one cluster. An example of this priority generation is shown in Figure 4.6. Looking at Figure 4.6, at merge level 0.9, both specimens 3111 and 3112 are set to a priority of 0.1. This is different compared to the unknown prioritization, where only one specimen receives a finite priority. In this algorithm, finite priorities are calculated for all specimens in the tree. Once all specimens have finite priorities, the specimens are sorted in decreasing order. The final priority list for Figure 4.6 is 3116, 3115, 3114, 3113, 3112, then 3111.

To eliminate specimen repetition, we generate individual trees following the level progression seen in Table 4.1. First, a tree using only the specimens with completely unknown identifications is created. These specimens are given the highest priority. Next we generate trees for each order identification, with unknown genus and species, or level 2 in Table 4.1. Each tree is prioritized separately. For example, benthic forams and planktic forams would have separate trees and be prioritized separately. Multiple trees are generated for the same reason as in identification generation, where merge levels may change. After all the priority lists are generated for level 2 in Table 4.1, these lists are combined together because no single tree should be above the other. If specimens in different trees are given the same priority level, they should all have the same effect on their respective subtrees so they should all be identified before specimens with a lower priority level. Similarly, priority is generated for specimens with known order and genus, but unknown species. At the end, the priority list for fully identified specimens is included.

The ideal priority algorithm was covered before the developed algorithm for un-

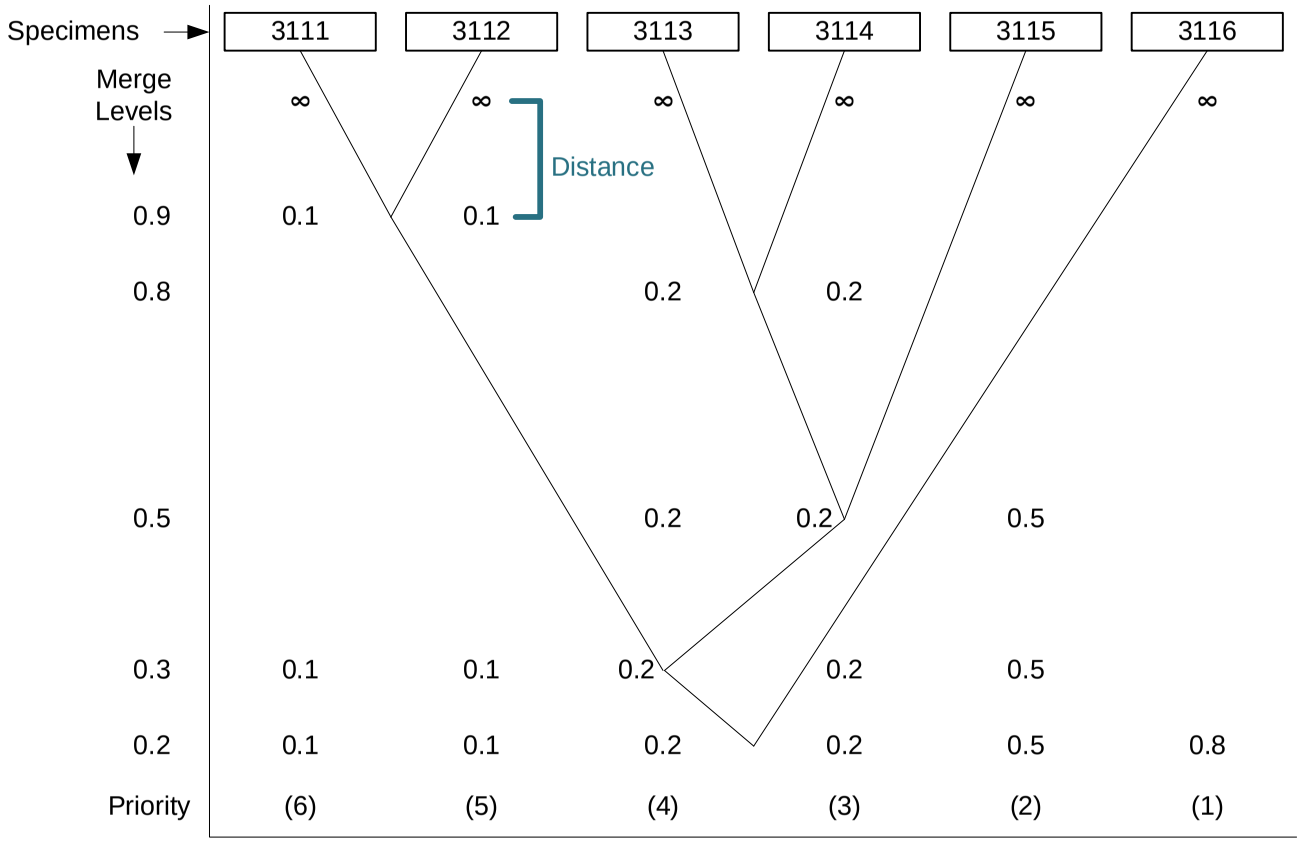


Figure 4.6: Fully known specimens also have a priority generated. This prioritization is similar to the prioritization of unknowns. However, all unprioritized specimens have a new priority calculated at each merge level.

known specimens, known specimens, and multiple trees were described. The goal for priority generation in the DHI algorithm is to assist with algorithm predictability and throughput, while ensuring reliable identification propagation (able to quickly increase the average dataset confidence). The DHI algorithm can be separated into identification propagation and priority generation parts, with both parts based on results from a tree generated using AHC. Together, these parts address the application requirements to quickly identify a high quantity of specimens with high quality identifications, and the additional predictability requirement for the dynamic learning inherent to crowdsourcing.

4.2 Results

In order to validate and verify algorithm behaviour and significance, we compared DHI to KNN. KNN is used because of the similarities it has to our approach through the propagation of identifications using a semi-automated algorithm and majority propagation [84, 85]. The simplicity of the KNN algorithm also makes it easier to modify to be used as a benchmark for hierarchical identification propagation. Tests were run using 238 specimens with Particle-Based Identifications (PBIs). As seen during the validation approach for the CASSIE 1 prototype, Image-Based Identifications (IBIs) can achieve correct genus 81% of the time and correct species 47% of the time [66]. Discrimination is difficult due to the loss of information when physical specimens are represented by simple images. As we now use digital representations that include both illumination and depth information, we expect correct rates for species to improve. Because we have yet to obtain an identified dataset of specimens using the Virtual Reflected-Light Microscopy (VRLM) applet, algorithm analysis uses previous PBIs only. As crowdsourcing is dynamic, we must evaluate how the algorithms behave as direct identifications are obtained over time. To validate the DHI algorithm, we examine the importance of identification sequence, compare DHI results to the best KNN results, examine the contribution of the priority algorithm, and examine the behaviour of the DHI algorithm at different thresholds and confidence levels.

To examine the role sequence plays on identification performance, KNN was run 40 times with a randomly generated sequence. To assess algorithm performance, we compare correct rates, indicating the percentage of specimens receiving the correct identification, and showing the thoroughness of generated identifications. We also compare incorrect rates, indicating the number of specimens receiving incorrect identifications, and showing the reliability of generated identifications. Unknown

identifications need not be counted when these two metrics are used. These metrics are compared to relative effort, which is the percentage of the dataset with direct identifications (a function of time). Crowdsourcing dynamics are simulated by repeatedly running the DHI and KNN algorithms using our dataset of 238 specimens incrementally, increasing the number of direct identifications one by one until the full dataset uses direct identifications, and no further identifications are propagated.

Figure 4.7 shows the maximum and minimum correct and incorrect rates versus relative effort when KNN is run multiple times. From the KNN results, a gap can be seen between the minimum and maximum values. This gap shows that identification sequence has an impact on algorithm performance. As crowdsourcing identifications are provided in an unpredictable manner, controlling the variability caused by sequence is important in a dynamic algorithm.

After confirming the importance of sequence, the benefits of the DHI algorithm can be accurately examined. The DHI algorithm is made of two main components, namely identification propagation and priority generation. Identification propagation can be justified because we want to increase the impact of one identification by allowing it to propagate to other unidentified specimens. The importance of the priority component of the DHI algorithm is also confirmed due to the variability seen in KNN performance. Next, we compare DHI and KNN performance. Looking at Figure 4.7, the DHI algorithm has good performance. The DHI graph is comparable to the best KNN performance with high correct and low incorrect rates. Using the DHI algorithm is better than KNN because the performance is reliable and predictable with the inclusion of the priority algorithm. However, DHI contains both identification and priority algorithms, so next we examine if the desired behavior is seen because of the priority algorithm alone.

To determine the amount of impact caused by the priority algorithm, KNN was run using the priority algorithm. In Figure 4.7, KNN using the priority algorithm does have the added benefit of predictable performance. However, it is unstable with correct and incorrect rates varying considerably as relative effort varies. This would be reasonable as the priority algorithm was designed to leverage the supervised algorithm performance. It can be concluded that the priority part of the DHI algorithm eliminates the variation in performance by providing a fixed sequence, while the identification part ensures reliable and stable identification results provided as quickly as possible.

While the DHI algorithm has been validated, we must also examine how algorithm behaviour varies as the threshold parameter varies. In the previous tests, the algorithms were run using a similarity threshold of 0.0. For the DHI algorithm, prop-

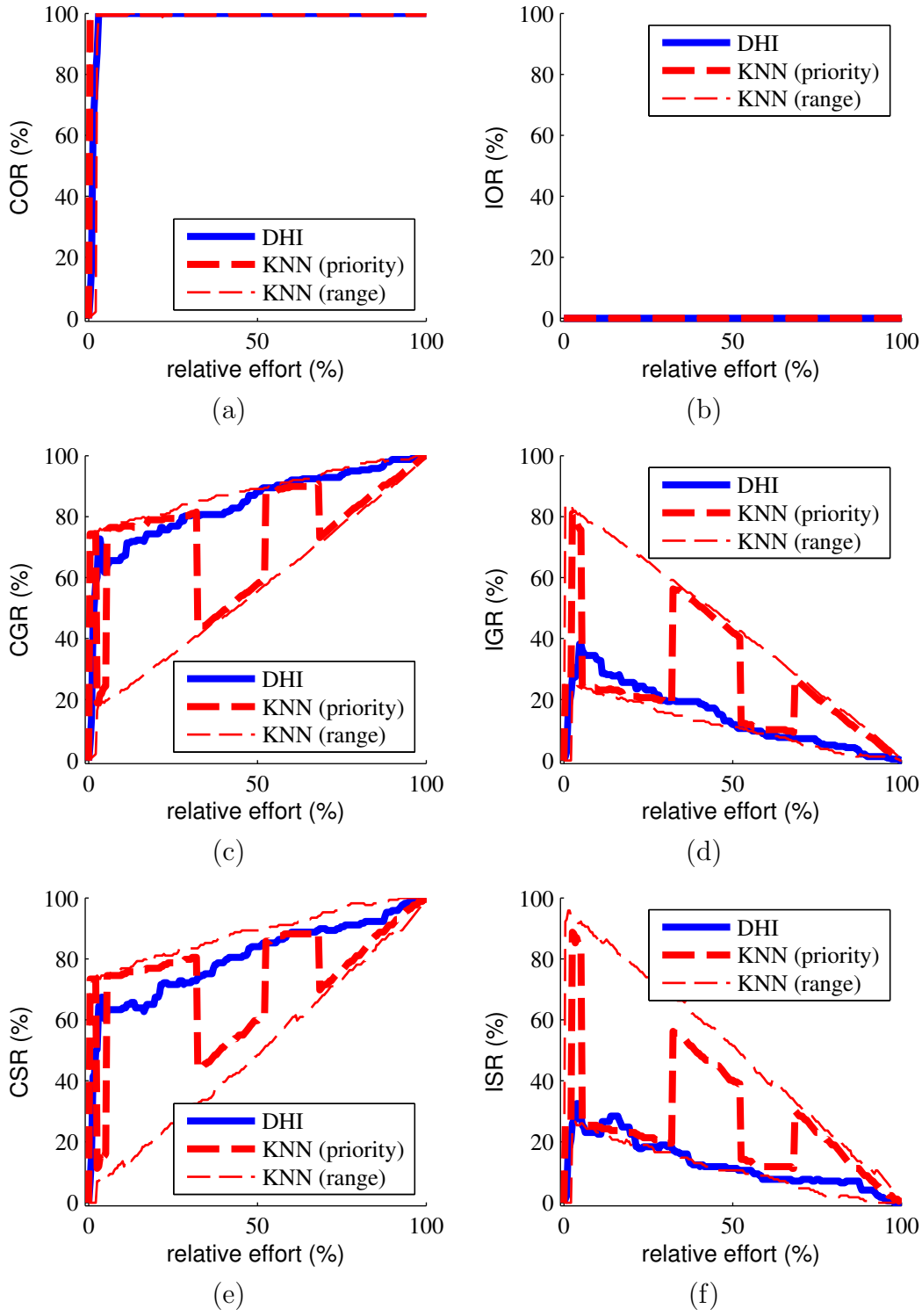


Figure 4.7: Identification results for 238 specimens were determined for three algorithms. Correct and incorrect rates for each taxonomic level are shown. Vertical axes show the percentage of correct or incorrect rates at the (a and b) order, (c and d) genus, and (e and f) species levels. Horizontal axes indicate the relative effort or percentage of specimens in the dataset with direct identifications.

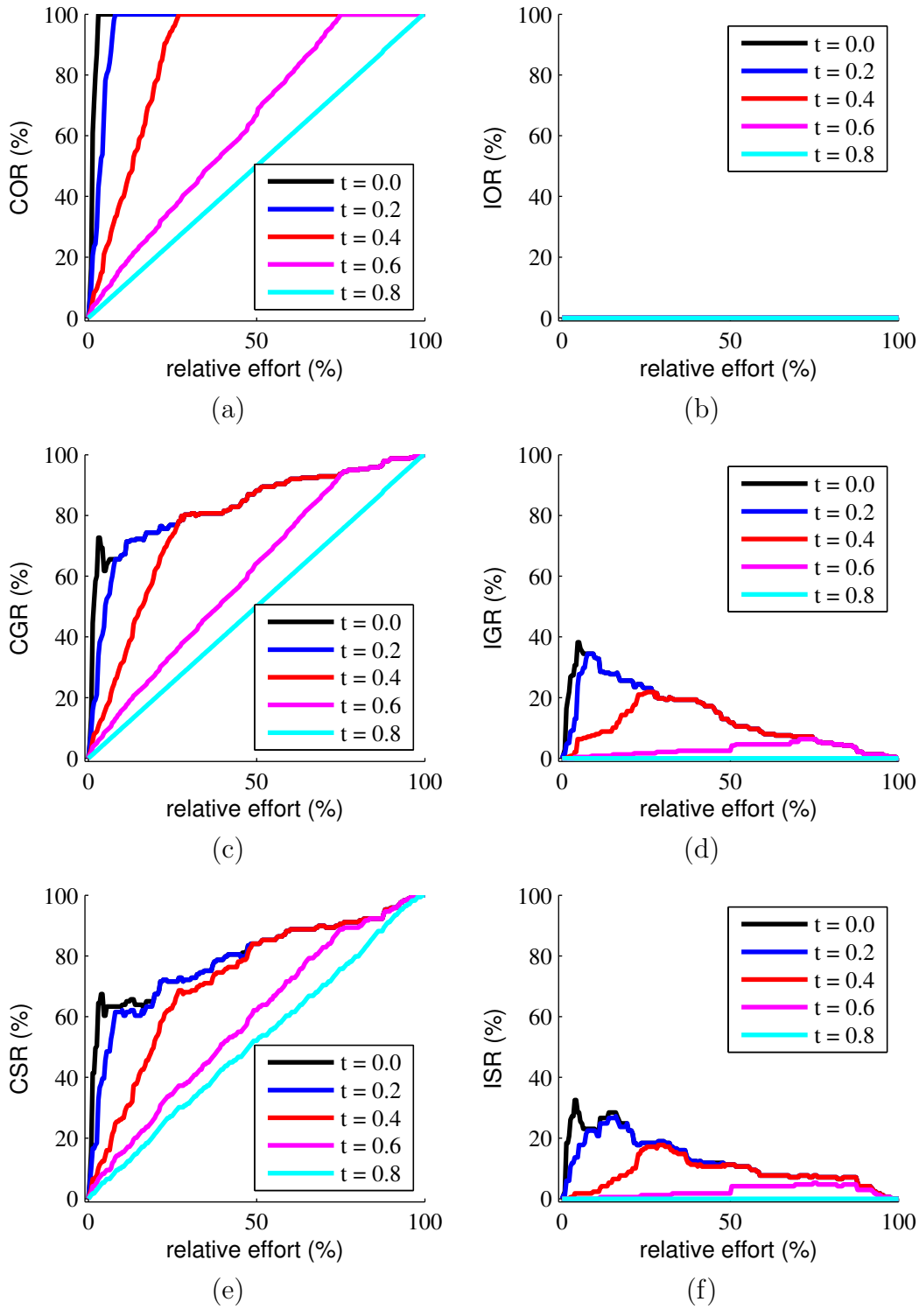


Figure 4.8: DHI correct and incorrect rates at varying thresholds for (a and b) order, (c and d) genus, and (e and f) species levels.

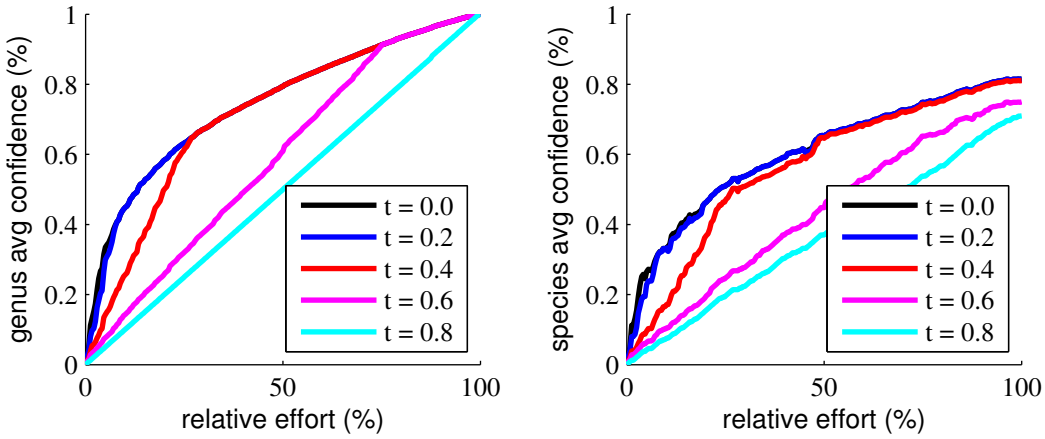


Figure 4.9: Average confidence for DHI identifications of 238 specimens at different thresholds for (a) genus and (b) species levels.

agation of identification is halted below the set threshold. Figure 4.8 shows DHI results as the threshold is varied. It should be noted that the diagonal line for correct rates means the direct identifications are the only identifications in the dataset. As direct identifications are not changed and will remain correct, they will always be seen as a diagonal in the correct rates. When no identifications are propagated, there are no incorrect identifications, so incorrect rates will remain at zero. Looking at Figure 4.8, correct rates are seen to approach the diagonal as threshold increases and incorrect rates approach a horizontal line at zero. This behaviour is expected. As the threshold increases, fewer identifications are propagated so more specimens contain unknown identifications and do not fall under correct or incorrect rates. To get the most out of this algorithm, it should be used with low threshold values where the most clusters are formed and the work reduced is greatest.

As identification confidence is provided, we can examine the confidence in the complete dataset over time (relative effort) to understand how reliability in the dataset varies with relative effort. Figure 4.9 shows the average confidence of the dataset as relative effort and threshold is varied. Examining these results show the average confidence in the dataset increases quickly at low thresholds. The average genus confidence approaches 1.0 with time, as expected, because all specimens end up with direct identifications of genus. The species results are similar to genus, but these results do not all approach 1.0 with time. The average confidence in the dataset is below 1.0 at 100% relative effort because relative effort is counting genus identifications whether or not species identification is provided.

To validate the behaviour of the DHI algorithm, identification sequence, algo-

rithm comparison, priority contribution, threshold contribution, and average confidence generation were examined. From these results, we are able to conclude that DHI behaves as desired, ensuring thorough, predictable, and reliable identification propagation while guiding volunteers to identify the full dataset quickly.

4.3 Discussion

DHI is a novel approach not used in other crowdsourcing projects. Introducing this algorithm to increase impact and throughput has many different implications. The dynamic approach to identification introduces new considerations in identification propagation and analysis, while allowing for hierarchical identification and code modularity that is customizable for many different applications.

The dynamic propagation of identification is a new approach to processing data obtained from crowdsourcing. Unlike most standard algorithms, DHI has no distinct processing stages: beginning (before clustering), middle (after clustering, before identification), and end (after identification). With DHI, cluster formation influence identification propagation. In addition to this, the direct identifications obtained using the priority algorithm feeds back and influences the indirect identifications. Note that the priority algorithm only suggests specimens for volunteers to identify first; they are not required to follow the priority sequence. While this does help ensure more predictable results, it does not guarantee results will be obtained in the sequence desired. With identifications being obtained and processed dynamically, and feedback affecting future identifications, the DHI algorithm does not have any set stages. By allowing more freedom in the way specimens are identified, we are able to generate complete results with an incomplete dataset.

Ability to dynamically identify specimens introduces new options in crowdsourcing analysis. In a dynamic approach, developers and researchers enjoy flexibility: analysis can begin at any point in the trial period; indirect identifications can be immediately generated for new specimens entered into the system; performance can be examined over time; and when database identifications reach a steady-state, results can be considered as ground truths. These new options enable researchers to better view, analyze, and use data generated from the system.

The hierarchical nature of the DHI algorithm is new to microfossil identification. Other identification approaches gather identifications of order, genus, and species essentially in one step. Here, DHI identifies taxon levels separately, generating more complete results. This type of hierarchical propagation of identification could be applied to many other applications, particularly in biology. DHI also works for ap-

plications without a hierarchy.

Modularity of the code design, which separates similarity generation, clustering, identification propagation, and priority generation enables the DHI algorithm to be customized and applied to many different applications. Changing any of these would allow the DHI algorithm to be customized to fit different applications. For example, adjusting how similarity is calculated can account for different digital representations of objects. This enables DHI to be applied in many different situations.

DHI introduces dynamic processing, dynamic analysis, hierarchical identification propagation, and customization for different applications. The adaptive and flexible nature of this algorithm makes it ideal for crowdsourcing projects and suitable for a wide range of applications.

4.4 Conclusion

To incorporate human-based computation into the Microfossil Quest system, algorithms were created to leverage the system for crowdsourcing. The main goal for the Microfossil Quest was to identify specimens in a dataset while ensuring thoroughness, reliability, predictability, and throughput. To meet these requirements, the DHI algorithm was developed and tested.

AHC was used as the basis for the DHI algorithm. Identification propagation and priority generation are both dependent on the trees developed from AHC. The tree is used to dynamically identify specimens, generating indirect identifications and linking them with a confidence level that helps indicate identification reliability. In addition to the identification generation, a priority algorithm was also developed to leverage human interaction. This prioritization ensures that we are able to obtain reliable direct and indirect identifications as quickly as possible for as many specimens in the dataset as possible. The priority algorithm also allows for more predictable algorithm behaviour.

When analyzing algorithm performance, we compared results with KNN as our benchmark because of its simplicity, making it easier to quickly modify the algorithm to propagate identifications in a hierarchical manner similar to DHI. KNN is also the ideal algorithm for comparison as it is a standard algorithm that is very similar to DHI because it is a semi-automated approach and uses majority voting to propagation identifications. Preliminary testing of DHI validated the ability for the identification algorithm to propagate reliable identifications comparable to the best KNN performance. The priority algorithm was also validated as it obtains predictable performance that is able to leverage the identification propagation algorithm. This

ensures reliable identifications, while increasing throughput. The algorithms have the most impact in reducing relative effort, while maintaining reliability, at lower thresholds. These tests confirm that the DHI algorithm helps meet the application requirements of the Microfossil Quest.

With the development and testing of the DHI algorithm, a thorough, reliable, predictable, and fast identification system for the Microfossil Quest was verified. This has strong implications for the area of crowdsourcing as this is the first attempt to dynamically generate complete dataset results. The approach will help obtain a large database of identified specimens that could be used in testing and validation of fully-automated microfossil identification solutions.

Chapter 5

Conclusion

Artificial Intelligence (AI) describes a computational system able to analyze, perceive, and respond in an appropriate manner to maximize the chance of success in a given context. A significant amount of research is focused on AI problems, covering a wide range of topics that include image identification. To facilitate study of the AI-complete topic of image identification, an application of sufficient importance must be chosen. Microfossil identification is such an application, where previous external work concerns rule-based and ANN-based approaches.

Microfossils, in particular marine microfossils, are important for study because of the rich information they have captured over millenia. With the order foraminifera, for example, researchers collect and identify forams for biostratigraphy to determine present day hydrocarbon accumulations and for geochemical analysis to determine prehistoric environmental conditions [15]. Due to their abundance in the world's oceans, and their relatively fast evolution, forams are ideal for these types of research [14, 15]. For best results, genus and species must be identified, which is a labor intensive process that is currently done manually. Large numbers of microfossils have already been collected through various ocean drilling programs, but only a small fraction of them have been identified. Introducing a fully-automated approach to microfossil identification would have a significant impact in assisting researchers, increasing the amount of information obtained, and facilitating the discovery of new applications.

Attempts to automate microfossil identification have been conducted since the late 1980s with the development of rule-based approaches. Rule-based systems include Fossil [33] and Visual Identification Expert System (VIDES) [34]. These approaches focused on assisting knowledgeable researchers or students with taxonomic identification through the refinement of a list of possible taxonomic identities. Previous rule-based approaches required users to examine microfossils under a microscope and

identify features manually, which does little to minimize the labour intensive nature of microfossil identification.

Fully-automated microfossil identification attempts have been conducted using Artificial Neural Network (ANN) approaches. The most well-known systems developed for fully-automated microfossil identification are CLASSIC [35], Computer Guided Nannofossil Identification System (COGNIS) Light [36], and Système de Reconnaissance Automatique de Coccolithes (SYRACO) [29, 37]. While more successful in reducing expert labour, ANN-based approaches met with limited success for different reasons. CLASSIC was able to successfully identify test specimens, but relied on difficult to obtain Scanning Electron Microscope (SEM) images that would also make any imaged microfossils unsuitable for geochemical analysis. COGNIS Light used optical images and was able to identify 93% of the *Florisphaera profunda* specimens being tested. However, COGNIS Light results are unreliable because an incorrect rate of 80% was also produced. SYRACO does not experience the issues seen in CLASSIC and COGNIS Light. Instead, the drawback with SYRACO is from the difficulty to justify generalization when using its *fat* ANN approach. When testing the 800 000 parameter system, SYRACO correctly identifies 91% of unseen faces after using *only* 200 training images. The difficulty in justifying such an approach, based on established ANN theory, inhibits the widespread use of the system due to, for some researchers, a lack of confidence in the approach.

While the importance of microfossil identification is established, rule-based and ANN-based approaches have limited success reducing labour and gaining popularity, while maintaining identification reliability. We propose the Microfossil Quest system, which is developed using a different approach to microfossil identification. Instead of a rule-based or ANN-based system, we approach automated microfossil identification incrementally, starting with a semi-automated system. With this partially automated system, the most difficult aspect of image identification—feature extraction and taxon identification—are performed using crowdsourcing. Crowdsourcing, or human-based computation in particular, is a new approach gaining popularity for research into AI-complete solutions. In this thesis, details on the general design, specific implementation, and preliminary results are given for the Microfossil Quest system. The contributions to knowledge made in the course of this work are described in Section 5.1. Plans for future work, including our long term objectives for the Microfossil Quest system, are described in Section 5.2.

5.1 Contributions

Previous identification research has been performed by others using rule-based identification, ANN-based identification, and crowdsourcing. Crowdsourcing projects are rapidly gaining popularity, but applying these approaches to microfossil identification has not been attempted before. A new system was developed of significance to software engineering, marine micropaleontology, and crowdsourcing. To contribute to software engineering, the evolutionary prototyping design life cycle is applied in Chapter 2 as a crowdsourcing case study. The human interaction component, which incorporates citizen cyberscience, contributes to marine microfossil identification, as explained in Chapter 3. Lastly, the development of a back-end for human-based computation, namely the dynamic hierarchical identification algorithm, is contributed in Chapter 4. These contributions are reviewed below.

5.1.1 Evolutionary Prototyping

A description of the design process for a specific crowdsourcing project is not given in the literature. We describe our design approach as a case study to aid other researchers. The benefits of the evolutionary prototyping design cycle, two previously-developed prototypes, and a third prototype design cycle were covered.

The long-term system we envision is being designed using an evolutionary prototyping design cycle. This approach allows for the reuse of previous prototypes to incrementally develop a complex system. We are developing a fully-automated identification system, which will benefit from this incremental approach.

Three prototypes have been developed at this stage. The first prototype, made by Kamal Ranaweera and called CASSIE 1, was designed to leave the most difficult aspect of microfossil identification, taxon identification, to experts. When evaluating the Computer-Aided System for Specimen Identification and Examination (CASSIE) 1, it was determined that illumination conditions in the images had a high impact on system performance, leading to CASSIE 2. CASSIE 2 was created by Adam Harrison and focused on obtaining better digital representations, addressing illumination variability with preliminary research into anaglyph representations. From CASSIE 2, a major bottleneck to further progress was identified in the difficulty of obtaining identifications for testing and validation of the system. This led us to a crowdsourcing approach in an attempt to address the difficulty in obtaining identified specimens. We created the Microfossil Quest prototype to convert from the previous computer-aided systems to a crowdsourcing system.

Requirements for the Microfossil Quest prototype are based on application-specific

and approach-specific goals. One application-specific requirement we imposed on the system was ensuring backward compatibility with the Microfossil Wiki data. The main application-specific requirement affecting system design is to reduce expert workload while ensuring identification reliability. For the new crowdsourcing approach, we introduce a new method to calibrate citizen data. In most citizen cyberscience systems, experts in the field of study are present on the research team to provide calibration data before trials begin. In the Microfossil Quest, experts are included in the system as volunteers providing identifications and not developers in the project. The identifications obtained by citizen volunteers are compared to expert identifications for calibration, giving us dynamic calibration data.

In the prototype modification stage, the Microfossil Quest system is separated into four components. Specimen acquisition was left relatively unchanged from CASSIE 2 with specimen dissemination only modified slightly. The human interaction component was entirely redone, creating a new website. The system also had major modification to the computation algorithms component, with the incorporation of two new algorithms: automatic prioritization and indirect identifications. All of the previously-existing components were completely redesigned for the new human-based computation approach with the main application requirement in mind.

Descriptions for the evolutionary prototyping design cycle, the three developed prototypes, and a high-level examination of the design cycle for the third prototype were outlined. From this high-level description of software development, more detailed descriptions were then given for the human interaction front-end and computational algorithms back-end.

5.1.2 Human Interaction

The human interaction front-end for the Microfossil Quest system was designed to add citizen cyberscience functionality along with replacing the old Microfossil Wiki website. The new website is the primary interface for the Microfossil Quest project and receives identifications, educates the public, and trains volunteers. The website itself is divided into five main sections: home, about, tutorial, system, and background.

The home page is where users can search the database and provide identifications. This is the page shown to users by default, making it easier for returning volunteers to immediately begin providing identifications. Unlike most other projects, users are able to search the database and provide identifications for any subset of specimens returned from the search. The search for specimens and identification are combined to give volunteers control over the specimens they identify. Volunteers wanting to

identify microfossils are able to search for any combination of features distinguishing specific types of specimens to identify, depending on their personal desire to identify what they know, or to practice identifying fields they do not know.

In the homepage, we include better digital representations than the Microfossil Wiki using the Virtual Reflected-Light Microscopy (VRLM) applet. The VRLM applet allows users to alter default illumination direction, type, and brightness for the current image. Users may also switch from non-anaglyph to anaglyph images that provide depth information. The resulting work on improved digital representations led to a publication [74], where the author's contribution involved the creation of the VRLM applet. No other citizen cyberscience project combines anaglyph, non-anaglyph, illumination, and depth information in its online digital representations. Other projects, like Stardust@home and Galaxy Zoo, use simple digital images of objects or, like Foldit, recreate models of objects from known features. These methods do not involve the automatic creation of models from physical objects.

The about section for the website is where new users are able to get a brief introduction to the system and our goals. This is also where volunteers are able to get answers to frequently asked questions and interact with each other in the forum.

The tutorial for new users is next. The tutorial is important to train citizens in microfossil identification. As this kind of project has not been done before, we focus on using images and text suitable for non-experts. There are many different microfossil taxonomies so the tutorial teaches features that may be found as opposed to describing each individual species. The current tutorial covers common microfossil orders, shell textures, chambers, apertures, and view sides.

The Microfossil Quest system description is provided in the system section. We believe allowing others to understand our goals and system design will encourage volunteers to participate and other research groups to create similar projects. For this reason, our website includes a high-level view of the complete system and our goals. While most other citizen cyberscience approaches give some information about the purpose and goals for the project, only Stardust@home included detailed information describing how the virtual microscope was made. Our approach describes all the modules in the system and how they interact.

Lastly, the background section covers elements important to the system. This is where users are able to learn why microfossils are important, the history of oceanography, how microfossils are obtained, and the forams used in this study. Crowdsourcing is the second major component of our project and descriptions about this field of research and other popular projects are also given.

Across the website, volunteers are able to learn and interact with the Microfossil

Quest system through the home, about, tutorial, system, and background sections. The human interaction component is seen by everyone and plays an important role in attracting and interacting with volunteers. This front-end to the system incorporates the citizen cyberscience role in the Microfossil Quest system. The next major component to the system is the human-based computation back-end using the computation algorithms.

5.1.3 Computation Algorithms

Our most significant contribution to human-based computation is our algorithms. The algorithms we developed are a combination of both supervised and unsupervised learning to leverage crowdsourcing. These are the first to dynamically identify the full dataset while users are entering identifications. This thesis described the algorithms developed for unsupervised, supervised, and dynamic learning with initial testing results included. Put together, these algorithms comprise a larger algorithm, which is called Dynamic Hierarchical Identification (DHI).

In order to cluster specimens, an unsupervised approach was taken. Trees were created using agglomerative hierarchical clustering to visualize and trace how clusters are formed. From these trees, indirect identifications are generated.

Using the trees, a supervised learning approach was used to propagate identifications as they are obtained. Indirect identifications allow for more identifications to be obtained quickly with reduced work effort. Other projects must wait for a given time period to obtain all identifications. The ability to dynamically generate indirect identifications allows for more freedom in the testing and analysis of system performance.

Closely tied to the propagation of identifications is the dynamic learning algorithm used for automatic prioritization. The Microfossil Quest system creates a priority list for the dataset. This ensures the dataset is quickly identified with both direct and indirect identifications. The prioritization is based on the assumption that volunteers have limited time to provide identifications. In order to make the most out of volunteer identifications, we suggest what specimens to identify to have the most impact on results. The indirect identification generation and automatic prioritization algorithms were both designed to increase throughput, thoroughness, and reliability for crowdsourcing, where identifications are provided dynamically.

Initial results from the DHI algorithm, made up of the above parts, shows we are able to quickly increase the thoroughness of identifications in the dataset by a comparable rate to the best performance of K-Nearest Neighbour (KNN). It was seen

that the sequence identifications are generated has an impact on performance, which could impact results in our dynamic approach. We have developed an identification algorithm for reliable performance, and an associated priority algorithm to leverage the identification algorithm in order to obtain reliable and predictable performance with high throughput. Results show that high correct rates, relatively low incorrect rates, and quickly increasing average confidence of the dataset is achieved at low confidence thresholds. This allows for the most impact on the dataset, while high confidence thresholds have little impact as few identifications are propagated.

The unsupervised, supervised, and dynamic learning parts of the DHI algorithm, along with preliminary results, were described. This algorithm plays an important role in our system, and is the first dynamic identification algorithm to be developed for crowdsourcing. The use of a dynamic algorithm for the back-end of the Microfossil Quest system is a new approach that would benefit other human-based computation and citizen cyberscience projects. The development of the Microfossil Quest system has contributed to research on microfossil identification, crowdsourcing system design, and leveraging for crowdsourcing. Our research shows considerations and approaches that can be taken when developing crowdsourcing projects.

5.2 Future Work

The work completed has developed the fundamental framework of the Microfossil Quest system. Most of the front-end interface has been implemented with the major functionality working correctly. In addition, a lot of the back-end computation has also been designed, tested, and verified to ensure expected functionality. As evolutionary prototyping will continually improve the Microfossil Quest system, more modifications to the system are being considered. Short-term goals for the Microfossil Quest system are described in Section 5.2.1. The long-term objectives for the system are explained in Section 5.2.2.

5.2.1 Short-Term Goals

Future development of the Microfossil Quest system should focus first on moving towards a prototype ready for testing with the public. This would require taking the framework and incorporating the general website features necessary to bring it to a testing stage for performance evaluation. This includes better integration of the identification and prioritization system with the website to create a more automated and continuous version of the Microfossil Quest system.

When the Microfossil Quest integration is completed, an alpha test of the system can be run. A usability analysis of the Microfossil Quest website must be conducted because of the importance placed on this interface to attract and interact with volunteers. On top of this assessment, the alpha test will also focus on obtaining expert identifications, verifying website information, and examining algorithm performance with a new dataset. Feedback from experts, specialists, and a few motivated citizens will be used to improve the system.

Once the base functionality and tutorial information of the Microfossil Quest system is verified, modifications will be made to incorporate specimen feature identification. By allowing for specimen feature identification, a wider range of volunteers are able to participate as features are easily seen in digital representations. An example of a feature would be the number of chambers seen in a specimen. This would make the system more appropriate for a wider audience, which is better for citizen cyberscience. Once the feature identification functionality is included, the full Microfossil Quest system can be tested. A trial will be run with system publicity to attract volunteers in order to determine volunteer interest and the feasibility of this approach for large-scale microfossil identification.

5.2.2 Long-Term Objectives

As indicated throughout this thesis, the Microfossil Quest system is being used as a step towards the bigger goal of a fully-automated system. Over the long-term, the system will be further refined, and used, in order to get closer to a fully-automated system. The long-term objectives for the Microfossil Quest system involve improvements for performance, automation, self-funding, and generalization.

Over the long term there are many possible areas that could be examined further to improve system performance or add to the Microfossil Quest system and its ability to give indirect identifications. Some areas in which the system could be improved include the use of different identification or prioritization algorithms. A more reliable similarity measure for agglomerative hierarchical clustering would also have an impact on results. Instead of image-based correlation coefficient, various other similarity metrics could be used, such as depth-map correlation coefficient.

To get closer to a fully-automated microfossil identification system, feature identifications from the citizen cyberscience part could be used by another component to provide computer-generated direct identifications of specimens. The features will enable other programs to do further processing and try to narrow down foram identification or even identify a foram based on features indicated by citizens. ANN-based

components could use citizen-provided feature identifications to narrow potential taxons, while using direct taxon identifications as part of the training set. Incorporating a rule-based component is also possible; this would allow for computer-generated identifications along with user-provided identifications. The benefit of both of these components is the reduced work for experts to obtain feature descriptions and more reliable feature descriptions given by humans. To facilitate some of these new components, modifications could be made to the VRLM applet to enable users to outline certain features that could later be used in processing. Having users indicate areas is not new as it has been used in both Stardust@home and HiRISE Clickworkers. However, with our multiple representations, a method would be needed to consolidate any location across all representations.

Once the system is sufficiently automated, digital representations along with their identifications could be sold or donated to aid further research or education. Physical specimens would be kept and archived. Once identifications are obtained, these physical specimens could be used for further study. In addition to this, researchers and industry could enter specimens into the system for a small fee. At this stage, we could have the system attempt to identify the specimens using stored identifications and indirect identification propagation, or a combination of the algorithms and citizen cyberscience. The amount charged for this use would go back towards system development and improvement. Such options would enable the system to be self-sufficient and assist with microfossil research, paleoenvironmental research, and oil exploration.

For system generalization, the types of microfossils entered into the database can be further expanded allowing for more types of microfossils to be incorporated. This would lead to a more complex tutorial that requires careful editing to prevent overwhelming new volunteers. With more microfossils added, a better database could be obtained to help anyone requiring microfossil identifications.

Improvement to performance, automation, self-funding, and generalization for the Microfossil Quest system will occur as research progresses. The Microfossil Quest system plays an important role in obtaining training and testing data. We hope to develop a system that can produce reliable results and provide a large database of identified specimens for further research and to develop better systems. Using the information provided from volunteers, automation research may be focused on creating better, more reliable, and accurate identification systems. Eventually, as more research is conducted, with the sufficient data for training and validation, we can develop a fully-automated system for microfossil identification, which has strong implications for image understanding, an AI-complete problem.

References

[1] Craig Gentry, Zulfikar Ramzan, and Stuart Stubblebine, "Secure distributed human computation," in *EC '05: Proceedings of the 6th ACM conference on Electronic commerce*, 2005, pp. 155–164.

[2] Stuart C. Shapiro, *Encyclopedia of Artificial Intelligence*, 2nd edition, 1992.

[3] Yuri Ostrovsky, Aaron Andalman, and Pawan Sinha, "Vision Following Extended Congenital Blindness," *Psychological Science*, vol. 17, no. 12, pp. 1009–1014, 2006.

[4] Colin Blakemore and Shelia Jennett, "blindness, recovery from," in *The Oxford Companion to the Body*. 2001.

[5] Wu Aimin, Xu De, Nie Zhaozheng, and Yang Xu, "Cognition Theory Based Performance Characterization in Computer Vision," *Advanced Concepts for Intelligent Vision Systems*, vol. 3708, pp. 210–218, 2005.

[6] Tommy W. S. Chow and M. K. M. Rahman, "A new image classification technique using tree-structured regional features," *Neurocomputing*, vol. 70, pp. 1040–1050, 2007.

[7] Roland Perko and Ales Leonardis, "A framework for visual-context-aware object detection in still images," *Computer Vision and Image Understanding*, vol. 114, pp. 700–711, March 2010.

[8] Xiping Luo, Jie Tian, and Yan Wu, "A Minutia Matching Algorithm in Fingerprint Verification," in *ICPR '00: Proceedings of the International Conference on Pattern Recognition*, 2000, pp. 833–836.

[9] Shahram Mohammadi and Ali Frajzadeh, "A Matching Algorithm of Minutiae for Real Time Fingerprint Identification System," in *ICPR '00: Proceedings of the International Conference on Pattern Recognition*, 2009, vol. 60, pp. 595–599.

[10] Jen Wu, Shan Juan Xie, Dong-Hun Seo, and Won Don Lee, "A New Approach for Classification of Fingerprint Image Quality," in *Cognitive Informatics, 2008. ICCI 2008. 7th IEEE International Conference on*, Aug 2008, pp. 375–383.

[11] Kevin J. Gaston and Mark A. O'Neill, "Automated Species Identification: Why Not?," *Philosophical Transactions: Biological Sciences*, vol. 359, no. 1444, pp. 655–667, April 2004, Taxonomy for the Twenty-First Century.

[12] National Science Foundation and The Regents, and Joint Oceanographic Institutions for Deep Earth Sampling, "Deep Sea Drilling Program," `http://www.deepseadrilling.org/about.htm`, last visited: May 2009.

[13] The American National Science Foundation and The Japanese Ministry of Education, Culture, Sports, Science and Technology, "Integrated Ocean Drilling Program," `http://www.iodp.org`, last visited: May 2009.

[14] Silvia Spezzaferri and Dorothee Spiegler, "Fossil planktic foraminifera (an overview)," *Palaontologische Zeitschrift*, vol. 79, no. 1, pp. 149–166, 2005.

[15] Howard A. Armstrong and Martin D. Brasier, *Microfossils*, Blackwell Publishing, 2005.

[16] University College London Micropalaeontology Unit, "Microfossil Image Recovery and Circulation for Learning and Education (MIRACLE)," `http://www.ucl.ac.uk/GeolSci/micropal/index.html`, last visited: May 2009.

[17] Jack R. Holt and Carlos A. Iudica, "Phylum Granuloreticulosa," `http://comenius.susqu.edu/bi/202/Taxa.htm` in Taxa of Life, last modified: March 2009.

[18] Christopher Taylor, "Rhizaria," `http://www.palaeos.com/Eukarya/Units/Rhizaria/Rhizaria.html#Acetosporea` in PALAEOS: The Trace of Life on Earth, last modified: 2004.

[19] J. Athersuch, F. Banner, A. Higgins, R. Howarth, and P. Swaby, "The application of expert systems to the identification and use of microfossils in the petroleum industry," *Mathematical Geology*, vol. 26, pp. 483–489, 1994.

[20] M. D. Simmons, W. A. Berggren, R. O. Koshkarly, B. J. O'Neill, R. W. Scott, and W. Ziegler, "Biostratigraphy and Geochronology in the 21st Century," *Paleo21*, 1997, `http://www.nhm.ac.uk/hosted_sites/paleonet/paleo21/biostrat.html`.

[21] D. Clay Kelly, Timothy J. Bralower, and James C. Zachos, "Evolutionary consequences of the latest Paleocene thermal maximum for tropical planktonic foraminifera," *Palaeogeography, Palaeoclimatology, Palaeocecology*, vol. 141, pp. 139–161, 1998.

[22] J. P. Kennett and L. D. Stott, "Abrupt deep-sea warmining, palaeoceanographic changes and benthic extinctions at the end of the palaeocene," *Nature*, vol. 353, pp. 225–229, September 1991.

[23] Jerome Landre and Frederic Truchetet, "Multiresolution Hierarchical Content-Based Image Retrieval of Paleontology Images," *Proceedings of SPIE*, vol. 5266, pp. 75–83, 2004.

[24] Stefan Fischer, Micael Binkert, and Horst Bunke, "Symmetry Based Indexing of Diatoms in an Image Database," in *Pattern Recognition, 2000. Proceedings. 15th International Conference*, 2000, vol. 2, pp. 895–898.

[25] Roberto Marmo, Sabrina Amodio, and Virginio Cantoni, "Microfossils shape classification using a set of width values," in *18th International Conference on Pattern Recognition (ICPR'06) - Volume 1*, 2006, pp. 691–694.

[26] Phil F. Culverhouse, Robert Williams, Mark Benfield, Per R. Flood, Anne F. Sell, Maria Grazia Mazzocchi, Isabella Buttino, and Mike Sieracki, "Automatic image analysis of plankton: future perspectives," *Marine Ecology Progress Series*, vol. 312, pp. 297–309, April 2006.

[27] Phil F. Culverhouse, R. G. Simpson, R. Ellis, J. A. Lindley, Robert Williams, T. Parsini, B. Reguera, I. Bravo, R. Zoppoli, G. Earnshaw, H. McCall, and G. Smith, "Automatic classification of field-collected dinoflagellates by artificial neural network," *Marine Ecology Progress Series*, vol. 139, no. 1–3, pp. 281–287, April 1996.

[28] John Cairns Jr., Silverio P. Almeida, and Hitoshi Fujii, "Automated Identification of Diatoms," *BioScience*, vol. 32, no. 2, pp. 98–102, April 1982.

[29] L. Beaufort and D. Dollfus, "Automatic recognition of coccoliths by dynamical neural networks," *Marine Micropaleontology*, vol. 51, no. 1–2, pp. 57–73, 2004.

[30] Alain Boucher, Pablo J. Hidalgo, Monique Thonnat, Jordina Belmonte, Carmen Galan, Pierre Bonton, and Régis Tomczak, "Development of a semi-automatic system for pollen recognition," *Aerobiologia*, vol. 18, pp. 195–201, 2002.

[31] Chun Chen, Emile Hendriks, Robert Duin, Johan Reiber, Pieter Hiemstra, Letty de Weger, and Berend Stoel, "Feasibility study on automated recognition of allergenic pollen: grass, birch and mugwort," *Aerobiologia*, vol. 22, pp. 275–284, 2006.

[32] Marc'Aurelio Ranzato, P. E. Taylor, J. M. House, R. C. Flagan, Yann LeCun, and Pietro Perona, "Automatic recognition of biological particles in microscopic images," *Pattern Recognition Letters*, vol. 28, pp. 32–39, 2007.

[33] D. R. Brough and I. F. Alexander, "The Fossil expert system," *Expert Systems*, vol. 3, no. 2, pp. 76–83, 1986.

[34] Peter Alan Swaby, "VIDES: An Expert System for Visually Identifying Microfossils," *IEEE Expert*, 1992.

[35] Shan Yu, Pierre Saint-Marc, Monique Thonnat, and Marc Berthod, "Feasibility study of automatic identification of planktic foraminifera by computer vision," *The Journal of Foraminiferal Research*, vol. 26, no. 2, pp. 113–123, April 1996.

[36] Jorg Bollmann, Patrick S. Quinn, Miguel Vela, Bernhard Brabec, Siegfried Brechner, Mara Y. Cortés, Heinz Hilbrecht, Daniela N. Schmidt, Ralpf Schiebel, and Hans R. Thierstein, *Automated Particle Analysis: Calcareous Microfossils*, pp. 229–252, Springer, 2004.

[37] D. Dollfus and L. Beaufort, "Fat neural network for recognition of position-normalised objects," *Neural Networks*, vol. 12, no. 3, pp. 553–560, 1999.

[38] Nigel Winser and Raghu Saxena, "Citizen scientists an untapped resource," *Science and Development Network*, April 2008, `http://www.scidev.net`.

[39] Science for citizens, "Science for citizens," `http://scienceforcitizens.net/`, last visited: Sept 2010.

[40] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players, "Predicting protein structures with a multiplayer online game," *Nature*, vol. 466, pp. 756–760, August 2010.

[41] National Audubon Society, Inc., "National Audubon Society," `http://www.audubon.org/`, last visited: May 2009.

[42] Geoffrey S. LeBaron, "The 108th Christmas Bird Count," *American Birds*, vol. 62, pp. 2–9, 2007–2008.

[43] Allison Childs Wells, "Log On to the 2nd Annual Great Backyard Bird Count," `http://www.birds.cornell.edu/Publications/Birdscope/Autumn1998/gbbc98124.htm`, Autumn 1998.

[44] National Audubon Society, Inc., Birds Studies Canada, and Cornell Lab of Ornithology, "Great Backyard Bird Count," `http://www.birdsource.org/gbbc/`, last visited: April 2011.

[45] Herbaria@home Team, "herbaria@home," `http://herbariaunited.org/atHome/`, last visited: June 2011.

[46] Luis von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum, "Improving Accessibility of the Web with a Computer Game," in *In ACM CHI Notes*, 2006, pp. 79–82.

[47] Luis von Ahn and Laura Dabbish, "Labeling images with a computer game," in *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 319–326.

[48] GWAP team, "Games with a Purpose," `http://www.gwap.com/gwap/`, last visited: May 2011.

[49] Eric Hand, "People Power: Networks of human minds are taking citizen science to a new level," *Nature*, vol. 466, pp. 685–687, August 2010.

[50] Andrew J. Westphal, Anna L. Butterworth, Matt Paul, Robert Lettieri, and Josh von Korff, "Stardust@home," `http://stardustathome.ssl.berkeley.edu/index.php`, last visited: May 2009.

[51] Andreq J. Westphal, Ronald K. Bastien, Anna L. Butterworth, Josh von Korff, David Anderson, Bryan Mendex, Rastika Prasad, Nicole Kelley, David Frank, Robert Lettieri, Zack Gainsforth, Christopher J. Snead, Jack L. Warren, Michael E. Zolensky, and 20 064 Stardust@home "dusters", "Search for Contemporary Interstellar Dust in the Stardust Collector," *Lunar and Planetary Science XXXVIII*, pp. 1–2, 2007.

[52] Aimee Whalen and Ron Baalke, "Stardust—NASA's Comet Sample Return Mission," `http://stardust.jpl.nasa.gov/tech/aerogel.html`, last visited: May 2009.

[53] The Planetary Society, "Projects: Stardust@home," http://www.planetary.org/programs/projects/innovative_technologies/stardustathome/, last visited: May 2011.

[54] Kate Land, Anže Slosar, Chris Lintott, Dan Andreescu, Steven Bamford, Phil Murray, Robert Nichol, M. Jordan Raddick, Kevin Schawinski, Alex Szalay, Daniel Thomas, and Jan Vandenberg, "Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey," 2008.

[55] Galaxy Zoo Team, "Galaxy Zoo 2," http://galaxyzoo.org, last visited: June 2011.

[56] Pietro Perona, "Vision of a Visipedia," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1526–1534, August 2010.

[57] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum, "reCAPTCHA: Human-Based Character Recognition via Web Security Measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[58] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford, "CAPTCHA: Using Hard AI Problems for Security," in *Proceedings of Eurocrypt*, 2003, pp. 294–311.

[59] Luis von Ahn, Mihir Kedia, and Manuel Blum, "Verbosity: a game for collecting common-sense facts," in *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 75–78.

[60] Robert Speer and Catherine Havasi, "Using Verbosity: Common Sense Data from Games with a Purpose," in *Twenty-Third International FLAIRS Conference*, 2010, pp. 104–109.

[61] ConceptNet Team, "Common Sense Computing Initiative," http://csc.media.mit.edu/conceptnet, last visited: May 2011.

[62] Shari Lawrence Pfleeger and Joanne M. Atlee, *Software Engineering: Theory and Practice*, Pearson Higher Education, 4 edition, 2010.

[63] Alan M. Davis, "Operational Prototyping: A New Development Approach," *IEEE Software*, vol. 9, no. 5, pp. 70–78, September 1992.

[64] Zooniverse Team, "Zooniverse—Real Science Online," http://www.zooniverse.org/home, last visited: June 2011.

[65] Kamal Ranaweera, Adam P. Harrison, Santo Bains, and Dileepan Joseph, "Feasibility of Computer-Aided Identification of Foraminiferal Tests," *Marine Micropaleontology*, vol. 72, pp. 66–75, June 2009.

[66] Kamal Ranaweera, Santo Bains, and Dileepan Joseph, "Analysis of Image-Based Classification of Foraminiferal Tests," *Marine Micropaleontology*, vol. 72, pp. 60–65, June 2009.

[67] Adam P. Harrison, "Computer Vision for Computer-Aided Microfossil Identification," M.S. thesis, University of Alberta, 2009, Department of Electrical and Computer Engineering.

[68] Virginia Gulick and Bob Kanefsky, "HiRISE Clickworkers," `http://clickworkers.arc.nasa.gov/hirise`, last visited: May 2009.

[69] M. Jordan Raddick, Georgia Bracey, Karen Carney, Geza Gyuk, Kirk Borne, John Wallin, and Suzanne Jacoby, "Citizen Science: Status and Research Directions for the Coming Decade," December 2010.

[70] Rick Bonney, Caren B. Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V. Rosenberg, and Jennifer Shirk, "Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy," *BioScience*, vol. 59, pp. 977–984, December 2009.

[71] Michelle Prysby and Paul Super, "Director's Guide to Best Practices Programming - Citizen Science (abridged)," 2007, Association of Nature Center Administrators.

[72] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg, "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, September 2008.

[73] Bernd Bruegge and Allen H. Dutoit, *Object-Oriented Software Engineering*, Pearson Education, Inc., 2 edition, 2004.

[74] Adam P. Harrison, Cindy M. Wong, and Dileepan Joseph, "Virtual Reflected-Light Microscopy," *Journal of Microscopy*, pp. 1–34, accepted: July 2010.

[75] Encyclopedia Britannica Online Academic Edition, "Foraminiferan," http://www.britannica.com/EBchecked/topic/212983/foraminiferan, last visited: February 2011.

[76] Australian Museum, "Microscopic marine creatures," http://australianmuseum.net.au/movie/Microscopic-marine-creatures/, last visited: February 2011.

[77] Encyclopedia Britannica Online Academic Edition, "Algae (biology)," http://www.britannica.com/EBchecked/topic/14828/algae/31714/Ecological-and-commercial-importance?anchor=ref958744, last visited: February 2011.

[78] Pamela J. W. Gore, "Historical Lab on Microfossils," http://facstaff.gpc.edu/~pgore/geology/historical_lab/microfossils.php, last visited: May 2009, Not available online anymore.

[79] Posted by ahnaf, "Terminology of Foraminiferal Test," http://foraminifera.net/foraminifera/terminology-of-foraminiferal-test.php, last modified: October 2008, last visited: March 30, 2011.

[80] Barun K. Sen Gupta, *Modern Foraminifera*, Kluwer Academic Publishers, Sept 1999.

[81] Geology.com, "Geologic Time Scale," http://geology.com/time.htm, last visited: May 2009.

[82] Richard Corfield, *Modern Foraminifera*, Washington, D.C. : Joseph Henry Press, April 2003.

[83] Anil K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, pp. 264–323, September 1999.

[84] Paul F. Whelan and Derek Molloy, *An Introduction to Machine Vision*, Oxford University Press, 2001.

[85] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions On Information Theory*, vol. 13, pp. 21–27, January 1967.

[86] Reynaldo J. Gil-García, José M. Badia-Contelles, and Aurora Pons-Porrata, "A General Framework for Agglomerative Hierarchical Clustering Algorithms," *Pattern Recognition, International Conference on*, vol. 2, pp. 569–572, August 2006.

[87] Danny Chen and Bin Xu, "Geometric Algorithms for Agglomerative Hierarchical Clustering," *Computing and Combinatorics*, vol. 2697, pp. 30–39, 2003.

[88] R. Ferzli and Lina J. Karam, "No-Reference Objective Wavelet Based Noise Immune Image Sharpness Metric," *IEEE International Conference*, vol. 1, pp. 11–14, September 2005.

[89] Nien-fan Zhang, Andras E. Vladar, Michael T. Postek, and Robert D. Larrabee, "Spectral density-based statistical measures for image sharpness," *Metrologia*, vol. 42, pp. 351–359, August 2005.

[90] Yong-Seok Choi and Sang-Joon Lee, "Three-dimensional volumetric measurement of red blood cell motion using digital holographic microscopy," *Applied Optics*, vol. 48, no. 16, pp. 2983–2990, June 2009.

[91] Lawrence Firestone, Kitty Cook, Kevin Culp, Neil Talsania, and Kendall Preston Jr., "Comparison of Autofocus Methods for Automated Microscopy," *Cytometry*, vol. 12, pp. 195–206, 1991.

Appendix A

Glossary

Anaglyph An image providing a 3D effect when viewed with red-cyan glasses.

Artificial intelligence The attempt to create intelligent machines and systems able to analyze and process information in order to achieve goals with comparable success to other intelligent beings such as humans.

Artificial neural network An adaptive system, modeled after biological neural networks, containing an interactive group of nodes that is able to change parameters and learn how to achieve particular results during the learning phase.

Citizen cyberscience A project educating and requesting volunteers to perform a task through the use of computers and the Internet in order to aid scientific research.

Citizen science A project educating and requesting volunteers to perform a task in order to aid scientific research.

Confidence level The amount of reliability in an indirect identification. The higher the confidence level, the more likely the cluster had high intra-specimen similarity when the identification was propagated and, therefore, the more likely the identification is correct.

Cluster A group of specimens.

Correct rate Number of specimens within a dataset that have received correct identifications, both direct and indirect, when compared to ground truth results.

Crowdsourcing The use of volunteers to perform a task.

Direct identification An identification provided by a human.

Distributed thinking A group of volunteers performing tasks to reach a common goal.

Double view When representations show two different sides of a specimen. The first view is the default side, visible to the camera when microfossils are initially placed under the microscope. The second view is the opposite side.

Dynamic hierarchical identification The algorithm developed for the Microfossil Quest system, made up of unsupervised, supervised, and dynamic learning parts.

Evolutionary prototyping Iteratively designing, modifying, and testing prototypes to create better systems.

Human-based computation A computer system that outsources tasks to humans in order to achieve a goal.

Image-based identification An identification provided by an expert or specialist after viewing an image or images.

Image understanding Interpreting objects, features, or regions within images to determine the details contained within the image.

Incorrect rate Number of specimens within s dataset that have received incorrect identifications when compared to ground truth results.

Indirect identification An identification generated by a computer.

K-means A method that partitions objects into k clusters by locating the most similar cluster template, as well as a method to calculate such templates.

K-nearest neighbour A method to identify objects by locating the k nearest objects that have already been identified.

Merge level The similarity score at which two clusters are combined.

Particle-based identification An identifications provided by an expert or specialist after viewing physical particles under a microscope.

Priority The rank in a list suggesting the sequence in which specimen identifications should be obtained to maximize the impact on the dataset.

Relative effort A ratio of the number of specimens with direct identifications to the total number of specimens in the dataset. This ratio may be incrementally increased until the full dataset has direct identifications.

Similarity A measurement to determine if foreground objects between a pair of digital representations are the same. We use correlation coefficients of image pairs, resulting in values from 0 to 1 with 1 representing perfect correlation and 0 representing no correlation.

Single view When representations show one side of a specimen, the default side that is facing the camera when specimens are initially placed under the microscope.

Threshold A number indicating of when to halt identification propagation. Any unknown specimens with similarity and/or merge levels above the threshold receive indirect identifications.

Tree A visual representation depicting the cluster relationships obtained by storing information on cluster formation during agglomerative hierarchical clustering.

Virtual reflected-light microscopy A system to create digital representations of opaque microscopic objects. This representation allows for anaglyph and non-anaglyph views while allowing for illumination control.

Appendix B

Special Cases

When conducting comparisons for the Dynamic Hierarchical Identification (DHI) algorithm there are some situations where specimens are tied and one must be chosen. A common occurance for these are when the leaves of the tree are involved. In order to choose a unique specimen, we refer to a study by Ranaweera *et al.* [65] where it was seen that image quality impacts identification accuracy. We assume that image quality is mainly affected by digital representation, in this case image sharpness. It is reasonable to assume that if images are blurry it is more difficult to see specimen characteristics and, therefore, to identify the specimen accurately. For this reason, image sharpness is used to resolve ties in priority generation as sharper images are easier to identify with a greater likelihood for identification accuracy. To compare image sharpness between specimens, a sharpness metric must be determined. The metric calculations, tests, and results are outlined.

In order to decide the best sharpness metric, we created our own sharpness metrics. By comparing a wide range of standard sharpness metrics—including variance, kurtosis, gradient, and Mendelsohn and Mayall's histogram methods [88–91]—it was decided that we would take a more general approach. Image sharpness computation was split into two steps, namely image processing and histogram processing. In the image processing phase, no processing, gradient, gradient squared, Laplacian, zero-mean Fourier transform, Fourier transform, zero-mean discrete cosine transform, and discrete cosine transform were used. Magnitudes were taken for vector or complex results. In the histogram processing phase, maximum, median, minimum, mean, standard deviation, kurtosis, quartile spread, mean square, and variance were used. Every pair, a combination of image processing and histogram processing, was used as a possible metric for sharpness comparison.

The testing procedure for the metrics involved running image and histogram processing pairs on all 238 specimens. Metric results for the 238 specimens were sorted

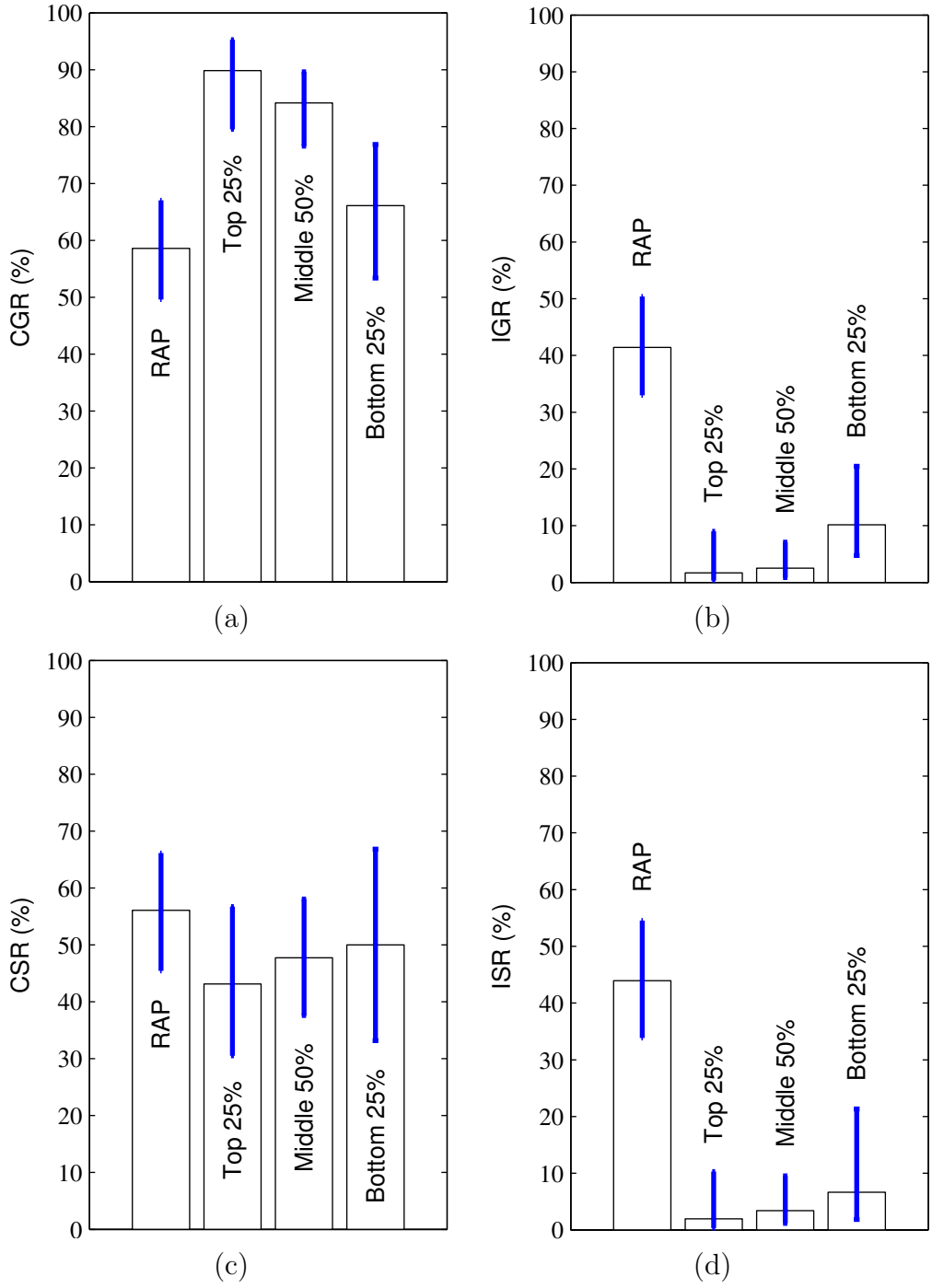


Figure B.1: Based on IBIs of 238 specimens, the best sharpness results were seen with Fourier transform image processing and maximum histogram processing: (a) CGR; (b) IGR; (c) CSR; and (d) ISR. 95% confidence intervals are shown, using the Wilson score method [66].

and split into three groups: the top 25%, the middle 50%, and the bottom 25%. The image identifications were compared to the ground truth in each subset and the correct and incorrect rates were calculated. When analyzing results, we looked for certain trends. An accurate metric should result in the top 25% of specimens with a high correct rate and a low incorrect rate, the middle 50% specimens with a lower number of correctly identified specimens and a slightly higher incorrect rate, leaving the bottom 25% specimens with the lowest correct rate and highest incorrect rate. Another assessment compared these to Random A Priori (RAP) results [66]: correct rates above the RAP and incorrect rates below the RAP are desired.

After processing was done, metric results were compared. Figure B.1 shows the Fourier maximum results, which had the best performance. Conceptually, the maximum absolute value of the discrete Fourier transform of an image indicates the amount of detail present in it. In Figure B.1(a), the desired decreasing CGR trend can be seen, with all values above the RAP. Similarly in Figure B.1(b), IGR shows the increasing trend as expected with all values significantly lower than the RAP. Figure B.1(d) shows the same desired trend with an increasing ISR, but Figure B.1(c) shows CSR is below the RAP with a slight increasing trend. This last result is attributed to the low number of species identifications that were obtained. Also, a low incorrect rate was deemed more of a priority. Overall, Fourier maximum results behaved as required, allowing for a fully-automatic metric able to resolve specimen ties.

In order to distinguish specimens in the case of ties, metric calculations were determined, testing was completed, and results were analyzed. From these results, it was seen that the Fourier maximum metric had the most desirable outcome, matching requirements. It is assumed from these results that sharper images are easier to identify accurately. This enables us to automatically generate image quality ratings and predict the relative quality of identifications for a specimen with a specific digital representation.

Appendix C

Pseudocode

This appendix gives the pseudocode for parts of the Dynamic Hierarchical Identification (DHI) algorithm described in Chapter 4. Explanations of these algorithm parts are given in the chapter.

Pseudocode for the unsupervised learning part is provided in Figure C.1. Unsupervised learning is explained in Section 4.1.1 using Figure 4.1. As similarity consolidation is also important, the pseudocode describing this is given in Figure C.2.

The supervised learning pseudocode is provided in Figure C.3. Supervised learning is the part of the DHI algorithm presented in Section 4.1.2 and Figure 4.3.

The dynamic learning part of the DHI algorithm is explained in Section 4.1.3 and incorporates slightly different behaviour for known and unknown specimens. The pseudocode in Figure C.4 is for the unknown specimen case, illustrated in Figure 4.5. The code for the known specimens is very similar to this.

The pseudocode for the three parts of the DHI algorithm is given to aid computer scientists with understanding. This appendix is meant to complement the explanations provided in Chapter 4.

```
Tree tree = new Tree(specimenTotal)
int iteration = specimenTotal;
do
    int maxSimilarity = tree.maxSimilarity;
    Vec clusterList = tree.findClusters(maxSimilarity);
    Cluster leftParent = clusterList(1);
    Cluster rightParent = clusterList(2);
    Cluster child = new Cluster(leftParent, rightParent, iteration);
    tree.add(child);
    iteration--;
while (child.specimenCount != specimenTotal)
```

Figure C.1: Unsupervised learning part of the DHI algorithm. The pseudocode describes how clusters are merged using AHC clustering.

```
left = leftParent; % Variable of type Cluster
right = rightParent; % Variable of type Cluster
Vec clusterList = leftParent.clusterList;
for (int i = iteration; i >= 1; i--)
    Cluster clusterNum = clusterList(i);
    if (clusterNum != leftParent && clusterNum != rightParent)
        int similarityLeft = leftParent.similarity(clusterNum);
        int similarityRight = rightParent.similarity(clusterNum);
        % Similarity is a variable of type Vec
        similarity.add(clusterNum, min(similarityLeft, similarityRight));
    end
end
```

Figure C.2: Constructor for the Cluster class that is used in the unsupervised learning part of the DHI algorithm. It is provided to describe how similarity pairs are consolidated when two clusters are merged.

```
while (!endOfTree)
    Cluster cluster = Tree.nextMerge;
    if (cluster.hasUnknown)
        Specimen specimen = cluster.hasUnknown;
        Taxon maxIdentification = cluster.maxIdentification;
        specimen.identification = maxIdentification;
        specimen.identificationConfidence = cluster.mergeLevel;
        cluster.update(specimen);
    end
end
```

Figure C.3: Supervised learning part of the DHI algorithm. The pseudocode illustrates how identifications are propagated in the tree. Processing begins at the first merge level after the leaves of the tree.

```
while (!endOfTree)
   Cluster cluster = Tree.nextMerge;
   % All specimens with priority of infinity
   Vec priority = cluster.infPriority;
   if (cluster.hasKnown)
      int numberOfInfinity = priority.size;
      for (int i = 1; i <= numberOfInfinity; i++)
         Specimen specimen = priority(i);
         specimen.updatePriority = 1 - cluster.mergeLevel;
         priority(i) = specimen;
      end
   else if (priority.size == 2)
      Specimen specimenOne = priority(1);
      Specimen specimenTwo = priority(2);
      if (sharpest(specimenOne, specimenTwo) == specimenOne)
         specimenTwo.updatePriority = 1 - cluster.mergeLevel;
      else
         specimenOne.updatePriority = 1 - cluster.mergeLevel;
      end
      priority(1) = specimenOne;
      priority(2) = specimenTwo;
   end
   cluster.updated(priority);
   Tree.update(cluster);
end
```

Figure C.4: Dynamic learning part of the DHI algorithm, for unknown specimens. This assumes all unknown specimens have an initial priority of infinity, and known specimens have an initial priority of negative infinity. The algorithm begins at the first merge level after the leaves of the tree.