**1. Vision – Global Names Discovery, Indexing and Reconciliation Services (GNDIRS)**

As scientific names have been almost universally used for 250 years, a names-based cyberinfrastructure will be an unavoidable part of the emerging Big Data version of biology [1]. This Development proposal will continue work that was started with the successful ABI Innovation grant [2]. We will develop **production level tools and services** that will discover names in digitized sources, will overcome **ambiguities in the meaning of names**; will **build an index of all names** and their uses, and through this **provide a comprehensive framework for access to biological data.**

The Global Names Architecture initiative in the US is extending the innovation grant in three complementary directions. The other two proposals are to develop the Global Name Usage Bank (Richard Pyle, Bishop Museum as Primary Investigator) and the Taxonomy Clearing House (PI David Patterson, Arizona State University). We plan tight coordination and integration between all the projects. By design Global Names Architecture is a modular system, with components developed independently, able to work in concert, but with the autonomy to be effective if other elements are not funded. Together, they dramatically enhance the diversity and usability of the system as a whole. Our vision of an interlinked system can be compared with the Domain Name System (DNS) used for resolving URLs to a particular computer in the global network. Our gateway will be used to resolve scientific names to a particular taxon across the spectrum of the 'Big Data' biology.

In this proposal we will address improvements in four areas:

**1.1 Scientific names discovery in texts and images.** A massive amount of data and literature contains information connected to scientific names [1]. Researchers interested in particular species need know where to find information about them. We will extend our prototype name discovery service to extract scientific names from terabytes of digitized data with high precision and recall (precision is the fraction of retrieved instances that are relevant, and recall is the fraction of relevant instances that are retrieved).

**1.2 Verification and disambiguation of scientific names.** Museum collection curators, biodiversity informaticians, scientists, biological data archivists, and historians often have collections with hundreds to billions of items that are linked to scientific names. To manage the information, it is crucial to know: which name strings are recognized scientific names, if they are spelled correctly, if the names are associated with the correct clade to overcome problems with homonyms, and what are the currently accepted names for outdated ones. It is useful to know what synonyms exist for the names, so that duplicates labeled with different names can be recognized. We will extend our prototype name reconciliation and resolution service to be able to go through millions of name strings and address these questions quickly and accurately. Reconciliation is the process that groups all name strings that are used for the same species, and resolution selects a currently accepted scientific name out of the reconciliation group.

**1.3 Index of name usages.** We will enhance our index of scientific names. It is similar to an index of terms at the end of a book, except much grander in scale. This index will include every name string that occurs in any data source collected by Global Names. The index will incorporate reconciliation, that is cross-reference lexical and nomenclatural variants of names. Such cross-referencing enhances the role of the index, allowing data sources to use Global Names to link out to a variety of other data sources. As a result, data providers can analyze what kind of species information they might be missing, or what name strings require review. The name indexing service will include reconciliation, user-customized resolution options, and will disambiguate homonyms.

**1.4 Name services as a gateway to biodiversity data.**  With our prototype developments we have confirmed that a combination of name discovery, reconciliation, and indexing creates an effective gateway to biodiversity data. GNUB (Global Names Usage Bank) [3] is designed as the source that will connect name strings to their nomenclatural foundations, synonymy, all nomenclatural events and literature references. GNUB’s technology is well developed and stable, has been designed for robust integration with our services. The supply of synonymy and taxonomic information will be enriched by the Taxonomy Clearing House project from Arizona State University. GNA accesses literature from CiteBank [4] created by Biodiversity Heritage Library.

**2. Background and Prerequisites**

In the second half of the eighteenth century, naturalists accepted the Linnaean approach to nomenclature. The emergence of a single unified system has obvious benefits, and for the last 250 years it has given us an unprecedented ability to point precisely to species of interest and has given experts around the world a common “biodiversity language”. Linnaean nomenclature is not perfect and some of the complications have become more and more pronounced [5] as we move into the Big Data world [1]. Linnaeus had been strongly influenced by the prevailing paradigms of creationism and essentialism and these influenced his nomenclatural principles. He assumed that the number of species was constant, boundaries between species were well defined and immutable, and the variability within species was stabilized and guarded by “essential features” defined by God. Species were assembled into a 5-rank hierarchy, and grouped based on essential features.

The binomial name serves dual purposes that can be the source of problems. One purpose is as an identifier, the other is as a mini-classification (placing a species within a genus). This dual functionality results in changes to the identifier when new taxonomic or phylogenetic information causes species to be placed into new genera.

Proper registration of new species names requires creation of an identifier as well as placement of it within a genus. For species that cannot be placed in a genus, various informal surrogate names have been used. For some clades of organisms, surrogate names vastly outnumber valid scientific names [6].

Linnaean classification ranks do not provide enough resolution to represent the trees built from phylogenetic information. Homotypic synonymy, heterotypic synonymy, alternative spellings and misspellings add a significant amount of ambiguity to the role of names as identifiers [1]. We approach the challenge of managing biodiversity information on the assumption that we must accommodate not only the Linnaean sense of species but also that of an increasing number of biologists who follow the evolutionary paradigm. From a phylogenetic point of view, the number of living species increases and decreases depending on environmental changes, random mutations and natural selection. Species can be placed into groups according to their phylogenetic distances, borders between species concepts are fuzzy, or non-existent, and taxonomy as a whole describes a dynamic process changing with time, and not a static landscape. A names-based infrastructure must be flexible, and allow for multiple co-existing interpretations of biodiversity

Out of the estimated 4-10 million species of eukaryotic organisms that currently populate the Earth, only about 2.3 million have been described [7, 8, 9]. The rest wait to be discovered or to be formalized from their surrogate name status. Some researchers call to dramatically speedup the process of species registration [10]. The Global Names Architecture must be designed for participation in this process.

As the amounts of scientific literature and data grow geometrically, biology has entered the scale of Big Data [11, 12]. It is not humanly possible now to follow all this information without the assistance of computers. Electronic publication of scientific papers is becoming ubiquitous and there is a significant effort in digitizing older information, making it available on Internet as raw scans and digital text via Optical Character Recognition (OCR) processes [13]. Our indexing services are being designed to intercept this flow of information, and by identifying the names, will provide organization to the heterogeneous content and make relevant content more discoverable.

**3. The Global Names Architecture – History and Progress to Date `**

**3.1. Background.** GNA was conceived [14, 15] in 2007 by Global Biodiversity Information Facility (GBIF) and Encyclopedia Of Life (EOL) to use names and associated information in expert sources to interconnect existing initiatives, enhancing them through their interactions as a virtual layer for managing biodiversity data distributed across the Internet [1]. GNA's conception involved an international consultative process with over 100 taxonomists, biodiversity informaticians, managers, and users of biodiversity data in the form of 13 Nomina meetings [15]. The initial GNA implementation was supported through an ABI Innovation grant [2] of which David Patterson was PI. Additional contributions are made by GBIF, EOL, Kew Gardens, PESI, and others. The purpose of the **ABI innovation award** was to investigate if the GNA vision was achievable. At the end of funding our combined team hasproduced a suite of proof of concept, prototype, and production-level services and infrastructure. Most notably -- an online registry for zoological nomenclature ZooBank [16], nomenclatural events repository GNUB [3], name finding service GNRD [17], name reconciliation and resolution service GN Resolver [18], name index GNI [19], taxonomic editor GNITE [20], and a citation repository CiteBank [4]. This ABI Development grant would bring funds to push name finding, indexing and reconciliation services from a prototype level of service to globally available production levels that will effectively serve the needs of Big Data Biology.

**3.2. GN1 Achievements.** **ZooBank** [16], is the on-line registry for zoological nomenclature for the International Commission for Zoological Nomenclature (ICZN) [21, 22]. ZooBank is a production-level system of services and a user interface that calls on the Global Names Usage Bank (GNUB) – a refined data environment for managing published statements about names. It is based on the Taxonomer model [3]. ZooBank has ushered in a modern era for nomenclature in which GNUB is now able to automatically capture information about new species as part of the publication process [23, 24]. **CiteBank** [4] operates as the article-level component of the Biodiversity Heritage Library [13]. The **Global Names Index** (GNI) [19] is the largest index of names with over 22 million name-strings found in data sources. These are the raw materials for GN reconciliation and resolution services that map alternative names to each other and correct names to those currently preferred by taxonomic sources [18]. Over 7 million of these names are annotated with taxonomic context so that homonyms (the same name used for different taxa) can be disambiguated. These environments as well as the **Interface for Taxonomic Editing** (GNITE) [20] are written in openly available Ruby code [25]. GNITE has a modern and powerful graphic user interface allowing the ingestion, display, drag-and-drop reassembly, and proof-of-concept merge of multiple taxonomies. It has already attracted 150 users in a short time span and is being tested for scale with a classification that initially had 800,000 entries. Through it, a phylogenetically inspired GN "**Union"** has been assembled, and was expanded to over 2 million entries by merging other major classifications to form a comprehensive taxonomic framework for OpenTree [26]. **PostBox** [27] is a **data entry portal** released to serve an EOL phylogenetic classification competition. GNRD, the **Global Names Recognition and Discovery** software [17], is used to identify and extract names from a diversity of data sources. It was delivered tothe Data Conservancy [28, 29] as part of its Feature ExtractionFramework to demonstrate the value of tools that indexes the heterogeneous content of data repositories such as DRYAD [30] by recognizing file types, unpacking the files, and identifying taxonomic names within the files. GNRD and GN Resolver provided new indexing services for the Biodiversity Heritage Library production site [31]. A full mirror of GNRD system was recently installed by Vibrant team to serve the EDIT and Scratchpads programs in Europe [32, 33]. GNRD and Resolver software are parts of the **NameSpotter** Chrome browser plug-in [34] that discovers names in documents and html pages in real time.

Over 30 GN applications and libraries are available through GitHub [25]. There have been over 150,000 downloads of our 9 Ruby libraries (gems) relating to Damerau-Levenshtein edit distance analysis, fuzzy matching algorithms, name parsing tools, and file conversion to the DWCA format (etc.). Our name parsing library is the most popular BioGem [35, 36] with 23,000 downloads. Our software is used by EOL, GBIF, Vibrant, iPlant, mx, Phylotastic, Arctos, iMarine, and WoRMS. We have worked closely with the iPlant team in developing the Taxonomic Name Resolution Service [37]. ZooBank presently holds over 100,000 nomenclatural acts and 40,000 publications, and is mirrored at multiple locations.

**4. Challenges and Implementation**

In planning the development of infrastructural services, we identify four levels of software development: **a)** **proof-of-concept** softwarethat confirms the feasibility of an idea, is quickly implemented, relatively cheap; **b)** **prototype services** that are made available on-line, but are not engineered to address all edge cases, nor work in all environments nor under all circumstances; **c) production level** services that are fast, scalable and stable, covering needs of 95% of possible use cases and are moderately costly to develop; and **d) mature services**,widely used, reliable workhorses - they are the most costly to develop.

We have used our previous ABI Innovation funding to create fully functional prototypes for the name discovery, indexing and reconciliation services. This ABI Development proposal targets three Global Name Architecture services: Names Finding [17], Names Reconciliation and Resolution [18], and Names Indexing [19]. This funding will allow us to extend existing services and move them to or beyond production level. For every objective in our vision for this work, we explain what is the current status for each service and the integrated system. Then we describe what needs to be done to reach production quality, and our approaches to meet the implementation challenges. The result of this work will be a suite of fast and reliable services with web-accessible user interfaces and APIs to solve problems associated with the handling of scientific names. Through the process of building the infrastructure we will create and maintain libraries of computer code [25] to empower other biodiversity informatics developers in creating innovative software.

**4.1** **Scientific names discovery in texts and images**

**4.1.1 Current Status:** The Global Names Recognition and Discovery (GNRD) project [17], takes plain text, web pages, PDF files, or pictures of a document in JPEG, GIF, PNG, or TIFF formats and returns back a list of names found in the source. It has a graphical user interface as well as an API. Additional optional information that can be returned includes the position of every occurrence of a name, a list of unique names, and language detection of the document. Output can be in html, xml, or json format. By combining name finding with name reconciliation services, found name strings can be verified in one step. For images, OCRed text can be returned upon request. In order to test our system, we applied it to the full corpus of Biodiversity Heritage library (105 000 titles, 40 million pages). Average throughput was about 1 million pages a day (name reconciliation was an independent step). The process is significantly slower for images and PDF files because of the additional overhead of transferring bigger files over the network and the processing costs of OCR. Currently, the system receives about 30 000 requests a day.

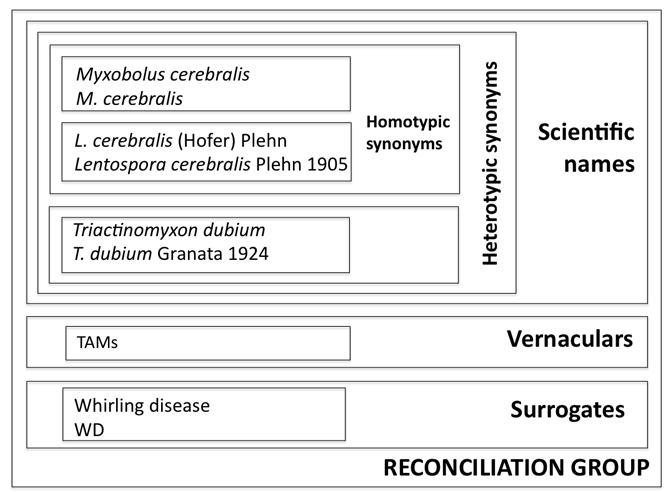
Name recognition and discovery is made possible because of two independent projects created at Marine Biological Laboratory. The TaxonFinder project [38], written by Patrick Leary, uses mostly heuristic algorithms that leverage “white lists” and “black lists” to recognize name strings. It is written in Perl and runs as a TCP/IP socket server that connects to GNRD. The second project, NetiNeti, uses a naïve Bayes classifier [39] for name recognition and is written in Python language by Lakshmi Manohar Akella during his tenure at MBL [40]. Unlike TaxonFinder, NetiNeti is able to discover names not previously known to us. NetiNeti runs as an http server to connect to GNRD. Both TaxonFinder and NetiNeti servers can run in multiple instances, using load balancer software to coordinate workload.

**4.1.2 Challenges: Speed, quality, complex names, abbreviated names.** At the moment the speed of name recognition and discovery by TaxonFinder and NetiNeti is limited by significant network overhead because components run on separate servers using different software platforms. The compilations of scientific names (“white list”) used by the projects [41, 42] are outdated and need to be updated to increase both precision and recall of name finding. Abbreviated names like ‘*D. melanogaster’* are used widely throughout scientific literature, and currently their discovery and resolution are poor. We set ourselves the task of dramatically improving performance that will expand such names to correct full species names. Name discovery algorithms do not find authorship information and perform poorly if there is authorship data in the middle of a name string. We will improve the NetiNeti algorithm in particular to accommodate lexical variants with authorship or infraspecies ranks.

**4.1.3 Implementation:**

**Speed and scalability**. We will re-implement TaxonFinder and NetiNeti as Ruby libraries to avoid network bottlenecks. We will run multiple instances in parallel using queuing libraries such as Resque [43]. **Quality.** We will update the “white” and “black” lists of TaxonFinder to add more names. This is an on-going iterative process – some scientific names are also common words in English and other languages so we will devise tests to ensure that both precision and recall increase over time. We will explore updating lists dynamically and use automated regression tests for quality control. NetiNeti currently generates many false positives especially with non-English texts. We already have a language detection system, which allows users to switch NetiNeti off for texts in languages where it performs badly. We will increase the accuracy of language detection, train the NetiNeti naïve Bayes classifier [39] to perform well with several popular languages (Italian, Spanish, French, German), and make NetiNeti adapt to the language of a given text. The current NetiNeti algorithm only recognizes canonical forms without ranks. With names like *Abacopteris simplex* Fée var. *trifoliata* Ching, or *Abacopteris simplex* var. *trifoliata,* only the species binomial component is recognized. We will change the algorithm to include rank and author recognition as well while insuring the process will be fast enough for production.

**Integration with name resolution.** Our experience is that false positives cannot be eliminated from the results of name discovery without diminishing recall significantly. We want to recover as many potential name strings as we reasonably can, so we developed an option of verifying found names by the GN name resolution service in order to identify and eliminate false positives. This is a relatively slow process, and we will explore accelerating it . We would store the original name string, the GUID of the name string it resolves to (if any), a resolution score based on a variety of parameters which indicates how good the match is, and a date. We can use the GUID to get more details about the cached name via GNI’s lexical and nomenclatural groups. This approach would dramatically increase the speed of name resolution for most users.

Figure 1. Name-string reconciliation involves mapping of lexical variants of scientific and vernacular names and surrogates for names; and appropriately flexibly and dynamically linking homotypic (objective) and heterotypic (subjective) synonyms.

**4.2 Verification and disambiguation of scientific names**

**4.2.1 Current Status:**

Reconciliation maps alternative names for the same species together (Figure 1). Given a name string, **reconciliation** allows to find corresponding scientific names that point to the same taxon. Reconciliation allows us to normalize names as different as “*Panagrellus redivivus”*, *“Chaos redivivum* Linnaeus, 1767”, *“Vibrio glutinis* Muller, 1783” (misspelled) and place them into the same group [44]. Closely related to reconciliation, the **resolution** process selects names that are “endorsed” by authoritative taxonomic sources -- *Pangagrellus redivivus* in this case.

The GN Resolver project [18] takes a list of name strings in question and attempts to match them to scientific names stored in the Global Names Index (GNI) database. Normally, it tries to find matches in all GNI data sources. If a data source states that the matched name is a synonym, the Resolver also returns the currently accepted name according to that data source. This tool is critical to addressing problems of synonymy, homonymy, and human errors accumulated over 2.5 centuries of binomial nomenclature history. Name resolution and reconciliation is a complex process and goes through six distinct stages (Table 1). First is the attempt to find exact match of a name string (fast), second is to find an exact match of canonical form (slow because it requires parsing of the input name strings), third is to attempt fuzzy matching of the name (slow, as it requires fuzzy matching algorithm). Fourth is the attempt to find an exact match of unmatched infraspecies names at the species level (fast), fifth is the fuzzy match of the same (slow), and finally the system seeks exact matches of unmatched species and infraspecies at the level of genus (fast). The resulting match score is heuristic and is expressed in a number from 0 to 1, where 1 is absolute certainty that the string is a name, 0 is certainty the string is not a name known to us, 0.5 is indecision (starting point).

Table 1. Six match types of name strings reconciliation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Match type | Name string | Name string to match | Matched by |
| 1 | Exact match | *Puma concolor* (Linnaeus, 1771) | *Puma concolor* (Linnaeus, 1771) | *Puma concolor* (Linnaeus, 1771) |
| 2 | Canonical form  exact match | *Puma concolor* (Linn.) | *Puma concolor* (Linnaeus, 1771) | *Puma concolor* |
| 3 | Canonical form  fuzzy match | *Puma cancolor* (Linn.) | *Puma concolor* (Linnaeus, 1771) | *Puma c[ao]ncolor* |
| 4 | Partial species  exact match | *Puma concolor cabreaae* | *Puma concolor* (Linnaeus, 1771) | *Puma concolor* |
| 5 | Partial species  fuzzy match | *Puma cancolor cabrear* | *Puma concolor* (Linnaeus, 1771) | *Puma c[ao]ncolor* |
| 6 | Partial genus  exact match | *Puma yagouaroundi* (Geoffroy, 1803) | *Puma* | *Puma* |

The largest project the system was tested against so far was finding names in full corpus of Biodiversity Heritage library. The average throughput was 600 000 names a day. Current usage of the system is about   
40 000 requests a day.

Several original algorithms are at the core of the reconciliation: Name parsing (atomizing) algorithm developed by Dmitry Mozzherin allows to find semantic meaning for different parts of a scientific name [45]. For example it determines that in Homo sapiens Linnaeus, 1758 ‘Homo’ represents genus string, ‘sapiens’ represents species epithet, ‘Linnaeus’ is author, and ‘1758’ is a year. It also finds a “canonical form” of the name – “Homo sapiens”. Canonical form normalization is crucial for reconciling names. Fuzzy name matching algorithm TaxaMatch was developed by Tony Rees [46]. It uses Damerau-Levenshtein edit distance algorithm [47] and fine-tuned heuristic rules to determine if two canonical forms are alternative spellings of the same name. In addition we use author matching algorithm developed by Patrick Leary.

**4.2.2 Challenges: Speed, accuracy, data harvesting**

**Automatic harvests of data.** We have adopted theDarwinCore Archives (DwCA) [48] format for data exchange between parts of Global Names Architecture. DwCA is also used to import name strings from external data sources into GNI database. Currently, harvesting of the names is not completely automated as we need often to map data from source to DWCA and to validate the results. We are now targeting automated updating processes that do not require human intervention.

**Speed, accuracy.** Parsing and fuzzy matching are the most expensive algorithms and create bottlenecks for the whole process. Improvements in the fuzzy matching algorithm, optimizing it for OCR errors, would significantly increase precision and recall for scanned texts. For high quality reconciliation, we need to know original (protolog) name, the chain of nomenclatural events, have access to corresponding literature, and to know detailed synonymy information.

**Homonyms resolution to a correct clade.** About 15% of the 463,000 compiled generic names are homonyms [49] where a name-string is used for more than one taxon. Homonyms confound federation of distributed data because information on unrelated taxa may be brought together, or because species with homonymic generic names have an uncertain identity. GNA collaborator Lafollette analyzed 532 titles from BHL (pers. comm.) that contained 350 pyramidellid (mollusk) generic names, yet only 35% were clearly about mollusks – the remainder concerned taxa made visible because the generic names were homonyms. There are many fewer homonymic species [49]. The problem is not trivial; as many as 3,000,000 names in GNI may be taxonomically ambiguous because of the generic part of the binomial are homonyms.

**4.2.3 Implementation:**

**Data and metadata acquisition.** Of two exchange formats that we use in the Global Names Architecture (Taxon Concept Schema [50] and Darwin Core Archive [48]), the latter has an advantage of being faster to transfer, and easier for data providers to create. We developed a Ruby software library called dwc-archive which allows efficient reading and writing of DwCA files [51]. Our dwca-hunter application downloads publicly accessible data in various formats (custom xml schemas, relational database dumps etc.) with conversion to DwCA [52]. Using these two programs we have been able to ingest scientific names, vernacular names, and classifications from ~70 different data sources, some of them more than 10 million records. Ingestion of very large sources slows the system down dramatically. We will make data acquisition an automatic recurrent process so that we can ingest large scale data without impacting overall system performance. Through a Nomina workshop, we have established that no IP rights apply to names and compilations, and will work to ensure that the names we collect will be accessible for interested parties, for example by other scientific name reconciliation projects like iPlant’s TNRS [53] or iMarine [54].

**Speed.** The two bottlenecks in this process are name string parsing and fuzzy matching. Name parsing is implemented as a Ruby library called biodiversity [45] and is based on a Parsing Expression Grammar [55] library treetop [56]. Biodiversity is our second most popular library and it is currently the most downloaded bio-library for Ruby [35, 36]. To speedup name parsing we will build a service that we will call insta-parser. It will be an open-source service that will cache parsed information for all name strings it encounters, and it will reparse them with every new release of the biodiversity library. Direct connection to the insta-parse database will significantly increase speed of the parsing activity.

For fuzzy matching, our system currently uses Damerau-Levenshtein algorithm [57], and it slows down in proportion to the product of the lengths of compared strings (O(N\*M) in big O notation where N and M are lengths of the strings)). We have already implemented the Damerau-Levenshtein algorithm in compiled computer language [58]. We consider to go further and implement the algorithm in assembly language to take advantage of MMX/SSE instructions of Intel processors [59] and to do several arithmetic operations in parallel during one processing cycle. Another approach that we will explore to increase speed is to use a generalized suffix tree algorithm in which performance is proportional to the length of name strings (O(N) in big O notation) [60]. We will investigate different options and pick one or more that are most effective.

**Quality**. Our parsing algorithm [45] continues to evolve constantly based on feedback from the user base as well as annual two week long concentrated improvement hackathon. The parsing library is currently at its third major version, and in its fifth year of development. We will continue further improvement of the library**. Quality of resolution and reconciliation** depends heavily on the amount of data available. Currently we operate with approximately 22 million name strings. Adding new data sources will help to resolve more names and with better accuracy. The Taxonomic Clearing House sister project, which is submitted through Arizona State University, plans to develop repository for taxonomies from various data sources. The GN project will delegate name and synonymy harvesting to them if they are funded. **Original names (protolog) and synonymy information, connection to literature** will be added to the Global Names Usage Bank (GNUB) [3]. There are already 500 000 protologs in GNUB and the number will grow further with additional funding. Both projects will significantly enhance quality of resolution in our service. **Abbreviated names** (like *H. sapiens*) are very common in the scientific literature and are hard to reconcile. We will address the problem that arises when several binomina in which the genus name starts with the same initial letter are present in the same source. We will reconcile them by searching for species epithets first and developing statistics for the best generic match for the binomen. The reconciliation **scoring system** thatwe have developed depends on a matching type for the canonical form, matching of authors, and the taxonomic context of a name. Taxonomic context is the common clade in a hierarchy to which at least 80% of submitted names belong. For example, if most names in a source are of vascular plants, context will allow our algorithm to resolve homonyms with a reasonable level of confidence. We will improve the scoring by taking account if a name is a known homonym, an ambiguous name (like Cancer or Acacia), if it is mentioned in curated, or uncurated data sources, how many data sources have the name. The largest compilation of homonyms with an estimated of over 95% of them is IRMNG produced by CSIRO in Australia[49]. We plan to create a classifier based on naïve Bayes algorithm, to decide if name should be considered “real”.

**Homonyms resolution.** There are several types of homonyms. During name resolution we mostly deal with homonyms of genera. Genera with the same spelling can happen in different codes. For example *Aotus* is the name of a genus of monkeys names in compliance with the zoological code (ICZN). It also refers to a member of a pea family names in compliance with the botanical code (ICN). Such homonyms are allowed and have to be dealt with. Also there are cases when the same name appears more than once in the same code. Such homonyms are not legal and are corrected by nomenclaturalists as they are discovered. Some of them persist. The occurrences of homonyms in legacy documents still have to be reconciled by our infrastructure. We will use homonym information from the open IRMNG database [49]. Such information will allow us to flag name strings as homonyms of a particular type, when available. It will also help to distinguish between homonyms and “chresonyms” – name strings where the associated authorship information is not relevant from the nomenclatural point of view. Species names that belong to homonymic genera will be reconciled by supplying known hierarchies from relevant data sources supplied through TCH. If the hierarchy is unknown, TCH - if funded - will integrate a FilteredPush annotation system to allow users to disambiguate homonyms and to mark chresonyms as such.

**Integration with Global Names Usage Bank.** Integration with GNUB provides a path of access to factual sequence of nomenclatural events, synonymy, literature information, and original descriptions of taxa (protologs). The Taxonomy Clearing House project (if funded) will supply synonymy and taxonomic information to GNUB, and hence will be automatically incorporated with our services. Integration with GNUB will happen in both directions. GNUB already uses a local Global Names Index database for fast access to GUIDs, lexical variants, and lexical groups. GNUB will be included as a data source for the Global Names Index and will be used for name strings reconciliations as any other data source. If a name string matches a scientific name from GNUB with allowance for homonyms through GNUB UUIDs we will be able to find the protolog name, currently accepted name, as well as other names from relevant nomenclatural events. Such information will be stored locally in GNI. For more detailed report about relevant literature and nomenclatural events we will send REST API calls to either a local mirror of GNUB, or to a remote GNUB instance. We will also maintain a UUID service which will resolve any UUID used in Global Names Architecture projects.

**4.3 Index of name usages**

**4.3.1 Current Status:**

Global Names Index project [19] allows to search by name string, genus, author, year, species epithet and shows which data sources have matching names, and gives links to every name usage. It is also powered by our first attempt to make lexical groups. It is the oldest service we have, and its development was mostly financed by Encyclopedia of Life and GBIF. The publically available code went through only minor updates recently and is becoming outdated. However GN Resolver uses very similar database structure and is planned to become the second major release of Global Names Index. Current version of GNI is an example of a stable code, which only crashed during last few years due to power or network outages. However from time to time GNI experiences slowdowns as a result of unusually high traffic. GNI usage is approximately 1500 page visits a day.

**4.3.2 Challenges: lexical groups, error correction.**

**Lexical and nomenclatural groups.** We can use resolution and reconciliation information to create a cross-referenced index among data sources that the Global Names Architecture is aware of. To speed up future resolution and reconciliation, we will create a service for fast crosslinking between data, allowing data sources to be enriched with information from other data sources. The creation of robust and comprehensive lexical and nomenclatural groups for all types of organisms is critical to this process, and is only being explored through Global Names.

**Error correction, annotation layer.** The aggregation of massive volumes of data from a variety of sources creates a problem of curation. People using the site will find mistakes and will want to correct them. Human input is a precious commodity and should be done only once for every mistake in a name string. Global Names services are not designed to fix errors originating in various data sources. However they would be able to edit a persistent annotation layer that is associated with the provided data. Such a layer, as is being developed by the FilteredPush team, would allow users to annotate and even correct misrepresentations. Creation of a globally available annotation/curation layer is a major challenge of Global Names Architecture and biodiversity in general. The TCH proposal includes requests for funds to integrate FilteredPush into Global Names.

**Surrogates.** Surrogates, or "dark taxa", are labels that *de facto* identify taxa but are not in the style of Linnaean Code-governed scientific names. They include labels like "*Coccinella* sp. 42" and, importantly, short gene sequences such as DNA barcodes or tags from high throughput sequencers [61, 62]. Submissions to GenBank without taxonomic names now exceed submissions with such names [6, 63].

In 2010, 87% of invertebrate records submitted to GenBank were surrogates [6], for example *Eptesicus* sp. A JLE-2010. If we are to avoid a profound epistemic split in diversity studies, it is critical for the names-based cyberinfrastructure to include surrogates, placing them in a taxonomically/phylogenetically meaningful name context so that data on all members of same lineages can be integrated [64, 65].

**Semantic world presence.** The emergence of the Linked Open Data Cloud (LODC) environment offers new semantic pathways through which data and metadata can be made available for unlimited analyses [66]. Biology is disproportionately highly active within LODC [67]. GNDIRS needs to be able to represent its data in a semantically friendly manner.

**4.3.3 Implementation**

**Index.** When a name string matches records from the Global Names Index database – a user gains access to GUIDs and outlinks pointing to more information from the relevant data sources. However pre-reconciling lexical variants of names and creating lexical groups would be very beneficial for users of our services. By including such processes, information will be retrieved very quickly using SQL queries against lexical groups tables. It is much faster to do this than sending a name string through the whole reconciliation and resolution procedure. Crosslinking between data sources allows us to show relationships, discover inconsistencies, misspellings, currently accepted names for outdated ones. For example if a data source comes from a museum which supplies names information from their holdings – it will be possible for us to generate update reports every time there is a change in nomenclatural information in any data sources they chose to follow. Two processes are involved. One is the creation of lexical groups, another is the creation of nomenclatural groups. Information held by GNUB will be used for nomenclatural grouping. Lexical grouping has to be performed first by reconciling all canonical forms held in GNI against each other. This process is already well developed and we have improved our algorithms to complete the task for all name strings known to GN in a matter of hours instead of days as with our first attempt at the outset of the GNI project. The information we receive is stored in a database. We then run full name strings which have matching candidates to see if their authorship information matches. This step is also well developed. The next step assembles the lexical groups. There is a danger of the inappropriate merger of non-related lexical groups because they have a common member. To avoid such situations, we will make sure that all members of a group match every other member and will create more than one group containing the same name string every time there is a mismatch.

**Vernacular names.** We collect vernacular names information when it is available through data sources. We will develop public interface to this data. Also we plan to connect to OpenUp initiative [68] to receive more exhaustive information about vernacular names.

**Surrogate Names.** We collect surrogate names from various sources. If we have information about their hierarchical position we are going to use this information to determine their relation to knows clades. Other treatments of surrogate names is out of the scope of this proposal. The Taxonomy Clearing House is the best candidate to deal with such items.

**Semantic presence.** We publish data from Global Names Index in RDF format already. We plan to work closely with Encyclopedia of Life to grow the presence of our data via LOD. Other semantic presence developments are out of the scope of this grant.

**4.4** **Name services as a gateway to biodiversity data**

**4.4.1 Current status:**

So far we had been building separate modular components of GN, and the interactions among them is still in infancy. We believe that a modular approach is beneficial as it enables users to mash them up in innovative and flexible ways. However some types of interactions are either necessary or extremely beneficial. For example ability to find names in texts and verify them in one step will streamline workflow and increase satisfaction for many users. We believe that future integration of name services from this grant with GNUB, synonymy information, CiteBank, and annotation layer will be powerful addition to existing functionality.

**4.4.2 Challenges:**

**Scope: Name-finding ~100 million pages, Resolution: ~100 million name strings.** Biodiversity related information connected to Linnaean nomenclature has been accumulated in ~500 million pages of scientific literature over 250 years, is in thousands of data formats and repositories, and maps to billions of preserved specimens with labels, to billions of triples in the Linked Open Data Cloud, as 200 million sequences in GenBank, and as 350 million records in GBIF [11, 69]. These data are being increasingly aggregated and new biodiversity data are being compiled by agencies such as DataOne, Data Conservancy, DRYAD, etc. The Global Names Architecture has been designed as a system that can process such a massive scope of data.

**Scale challenge: 100-300 requests per minute.** At the moment we can only estimate the scale of possible usage. The demand from Biodiversity Heritage Library alone requires that GN processes about 50 000 pages weekly just to keep abreast with newly scanned pages. We intend not only to process this information in one day, which equals 34 calls per minute to the name-finding and name-resolution services, but to be able to accommodate a 3-10 times bigger load at the end of the grant period. That is, we expect to respond to 100-300 calls per minute. While this does not meet our most aggressive projections, this target will satisfy most users, and will allow us to explore and try scaling procedures to make sure that we are able to scale according to the demand.

**Mirroring, stand-alone appliance.** There are two different use cases in replication of the GNDIRS services. One is to mirror data and code between participants of Global Names Architecture initiative. Another one is to allow large scale users, for example Biodiversity Heritage Library, iDigBio, Vibrant, Chinese Academy of Science to establish local instances so they can access services without having to contend with the lag problems of accessing services via internet. An elegant solution to both issues to create scripts which would allow to automatically produce full copy of the GNDIRS services on an empty, or pre-virtualized hardware. Such an approach would have to be self-sustainable with automatic upgrades of the code and data when there are updates on the central server.

**Reusability: code libraries, applications, services.** No system can be perfectly suited for all potential use cases and players. Our goal here is to break the software, especially algorithmic parts into stand-alone libraries, applications and services, and make them available for reuse by third parties.

**Sustainability challenge.** Infrastructure based on periodic funding is inherently unstable. This grant is a short-term solution for sustainability of the project. Long term sustainability of this project needs to be addressed in the future.

**4.4.3 Implementation:**

**Scope.** We strive to be able to scan through all available literature and other data sources that are available in a digital form. We also expect to be able to resolve all name strings found. We started to adjust to the scope when we reindexed all the literature collected in Biodiversity Heritage Library. We ran 105515 titles, 39 million pages through our name finding and name resolution services. The process took us 2 months, with name reconciliation being a limiting factor. Out of 27 million name strings found only 10 million had been verified as scientific names. The reconciled name strings in turn had been normalized to 2.6 million canonical forms, giving on average about 4 lexical variants per name. The majority of false positives came from texts written in Romance languages other than English (Italian, Spanish, German etc.). The BHL reindexing showed that our prototype can process approximately 20 million pages and 14 million name strings per month. Further improvements are possible, including delegating some of the work to other instances of the Global Names software. We will implement speed enhancements and upgrade our server base and will be able to go through up to 200-400 million pages and 100-300 million names per month using our site only. We think it should be sufficient for the next phase of resolution services to meet global need.

**Scale.** We have 2 goals in scalability – to be able to have near linear performance increase by adding new hardware for main site and for mirrors, and to be able to create stand-alone name server appliances with full functionality for users who want to avoid network lag and run services locally. We will move the services to a cloud-computing environment using OpenStack OS [70]. The hardware that we purchase is going to be integrated into the cloud and virtual machines for running name services and will be provisioned using Chef [71] – scripting platform for automated deployment and configuration of virtual services. We will have several instances of name discovery and name resolution servers running behind a load balancer. Name resolution will depend on the Global Name Index database. This database will have one master database and several slave databases. A master database will be responsible for all the harvesting. Further parallelization will be achieved by using several workers connected to a job queue for easily parallelized tasks, like name finding, and name resolution. We will also explore increasing parallelization of the processing. We will have an insta-parse server for local and public use to speedup the name string parsing process. We will automate the process of creating the GN appliance servers. Users of such appliances will have a choice to register them at our update service. If they will do so – their database will be automatically synchronized, their code base will also be updated with every formal release of the code via automated scripts.

**Mirroring.** We plan to create full publicly accessible mirrors of the service to increase scalability as well as decrease network lag for different regions of the globe. Our mirrors will use GN appliance technology, when it is ready for production. We plan to have full mirrors of the services running from other participants in Global Names Architecture Initiatives. We will have mirrors installed at Arizona State University, AZ, USA, Bishop Museum, HI, USA, Kew Royal Botanical Gardens UK, Chinese Academy of Science Beijing, USA. We also plan to encourage collaboration in development effort with the biodiversity groups abroad. Such global collaboration, would be one of the sustainability strategies.

**Reusability.** During development we have and will pay very close attention to the reusability of the code created. We work to solve every non-trivial computing problem we encounter as a stand-alone library. All our libraries are released under open source licenses (MIT and GPL v.2). We publish all the code using the public repository site, Github. As a result, all the changes made by us are available essentially in real time through the repositories. We write exhaustive acceptance and regression test suites to “document” library functionality and to avoid bug introduction. The APIs of each library are fully documented in English and available at the code repository, as well as being supplied with the library itself. Every change introduced in the library is run through a Continuous Integration server, often through several Ruby language versions. We pay full attention to users’ feedback, and carefully consider every patch submitted by users. So far we have incorporated all of the patches sent to us. To ease the upgrade burden for our user base we follow naming conventions and best practices of semantic versioning for our libraries. We also follow guidelines for Ruby language style and refactor our libraries when opportunities arise. All the libraries are easily installed using Ruby standards. So far we enjoy growing popularity of our developments, and up to date our libraries had been installed ~150 000 times.

**Sustainability.** Short term sustainability is going to be achieved via this grant. Also we will pursue following opportunities:. We will use funds available to other Global Names participants to have full mirrors of our services. We will continue to solicit funding from emerging opportunities. We will increase integration of our services with Encyclopedia of Life. We will encourage more active participation of Global Names Europe and Global Names China. We recently received a supplementary award for creating full mirror of our services at Chinese Academy of Science and plan to have it working, and transfer knowledge about maintaining it by the end of 2013.

**5. Overview of GNDIRS Production Time Line**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2.** Anticipated time line for project development tasks. Timing assumes a project starting date of 1 April 2014. Alternating column coloring represents project years, rather than calendar years. D=Development (including Testing); M=Maintenance; | | | | | | | | |
|  | **2014** | | |  | **2015** | |  | **2016** |
| **Development Tasks** | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 |
| **Scientific names discovery in texts and images** |  |  |  |  |  |  |  |  |
| TaxonFinder upgrade as Ruby library | D | M | M | M | M | M | M | M |
| NetiNeti upgrade as Ruby library |  |  |  |  | D | D | M | M |
| Scalability improvements |  |  |  |  |  | D | M | M |
| **Verification and disambiguation of scientific names** |  |  |  |  |  |  |  |  |
| Integration with Global Names Usage Bank | D | M | M | M | M | M | M | M |
| Reconciliation/Resolution improvements |  |  |  | D | D | D | D | M |
| Automatic Harvesting |  | D | M |  |  |  |  |  |
| Parsing algorithm | M | M | M | M | M | M | M | M |
| Fuzzy matching algorithm | D | D | M | M |  |  |  |  |
| **Index of name usages** |  |  |  |  |  |  |  |  |
| Lexical groups development | D | D | M | M | M | M | M | M |
| Migrating of GUI and API to new code base |  |  | D | D | M | D | D | M |
| Integration of nomeclatural groups |  |  |  |  | D | D | M | M |
| **Name services as a gateway to biodiversity data** |  |  |  |  |  |  |  |  |
| Scalability and performance tuning |  |  |  | D | D | M | M | M |
| Deployment, mirroring, syncronization scripts |  |  | D | M |  |  |  |  |
| Scripts to create Name Services appliance |  |  |  | D | M | M | M | M |

Before the last 3 months of the grant period we are planning to have all features we indicated to be developed. We are going to spend last 3 months concentrating on usability of the services, polishing the interface and functionality, making sure that we do accommodate accepted use cases as well as we can.

The very last month of the grant period will be considered a feature freeze. We will work on test coverage, stability issues, and automation of system administration – to make sure that the projects are able to run as reliably as electric power and hardware robustness allows.

**6. Qualifications of the Investigator and the Grantee Organization.**

Nathan Wilson, Principle Investigator, leads development of Encyclopedia of Life, one of the biggest biodiversity projects in the world.

Dmitry Mozzherin, Co-PI, has a Ph.D. in biology. He has been working as a professional programmer for 13 years, five and a half of which he worked on a large scale biodiversity projects. Over the last 5 years he wrote the majority of the code for the Global Name discovery, indexing and resolution projects. He is intimately familiar with the code from other parts of the projects. He also wrote 9 biodiversity-related libraries, one of them – Scientific Names Parser is the most downloaded biology library for the Ruby language.

**7. Impact of Activity on science, infrastructure, and education.**

**7.1 Science.** GNDIRS together with other Global Names Initiative is aimed to evolve into a pervasive layer, connecting data through the aspect of scientific names. This virtual layer with stable, robust, and quick services will promote the transition of biodiversity-based sciences into the Big Data world. Indexing services will increase visibility and awareness of new content that is indexed through GNUB and made accessible using RESTful services and semantic web. Significantly, reconciliation and resolution services will help to normalize the taxonomic terms that are critical to data integration. By interconnecting existing names-based data centers and services, GNA provides a unique solution that will accelerate discovery and access to data. The project has international buy-in; benefits from critical input and prototype development over 12 years, and links to data standards initiatives (e.g. GBIF and TDWG) and to other biodiversity data and infrastructure initiatives within the U.S and beyond.

**7.2 Infrastructure.**

We aim to promote 'names thinking' as a familiar part of established data management procedures, to increase the engagement of students with the project, and to position Global Names within the context of U.S. and international biodiversity information management projects. Primary beneficiaries of GN2 will be the research and museum communities tasked with producing and curating name-based information in the biological sciences. Nomenclaturalists will benefit from a paradigm-setting scientific names registration environment [72, 73]; taxonomists will have access to larger compilations and tools to build and semantically track comprehensive classifications; and biodiversity data managers will gain access to names reconciliation and resolution services as GN2 indexing services draw more of the long tail of science into daylight [11, 69]. API usage opens services for educators and general public via Encyclopedia of Life, Biodiversity Heritage Library, museums databases like Arctos. All code will be instantly available as open source to be used as libraries, or for learning how algorithms function. NSF and the community of research project PIs will benefit from services that provide automated taxonomic indexing of websites, documents, and databases as a component of Data Management Plans, thereby also allowing NSF to create a common names index to supporting projects. Stakeholder input – we will work with consultants in evaluating stakeholder needs and satisfaction, to develop and implement our sustainability plan.

**7.3 Education.**

Undergraduate and graduate students who use data for their education and research (for example Embryology project run by Arizona State University at MBL) will be invited to participate in the project. Biodiversity informatics training courses developed at MBL will show students how to integrate the services in their workflow, and provide an overview of biodiversity informatics. Our group will continue to closely cooperate with other members of Global Names initiative, to integrate our work more closely with EOL, as well as cooperate externally with Arctos, Vibrant, Plazi, iPlant, etc., and so will expand visibility and our user-base through improving performance and scalability.

**8. Results from the prior NSF support**

Nathan Wilson (PI) INSPIRE-1243575 “A Digital HPS Infrastructure for Understanding Biodiversity”, 9/12-8/14 $599,999. PIs J. Maienschein, N. Wilson, J. P. Collins, M. Laubichler. As part of this project Wilson led the effort to deliver the publicly accessible History of MBL and Embryo Project websites which use the History and Philosophy of Science Repository jointly developed at MBL and ASU. The next phase of this project includes cross-linking and data transfer between these sites and the Encyclopedia of Life.

Dmitry Mozzherin (Co-PI) N/A