BIOSPECIMENS AND THE INFORMATION LANDSCAPE FOR BIODEFENSE:

WORKSHOP REPORT

A Product of the Interagency Working Group on Scientific Collections

D. DIEULIIS, N. BAJEMA, N. WINSTEAD

JUNE 2019

Executive Summary

US biodefense and biosecurity rely on the ability to perform broad biosurveillance, to protect and secure biological agents of concern, as well as to diagnose and mitigate the potential consequences of the spread of global infectious diseases. Each of these abilities depends upon the collection and identification of biological samples, or, biospecimens. As genomic sequencing and synthesis tools continue to grow, the genomic information associated with biospecimens is expanding rapidly; the rapid convergence of the physical and digital worlds through digitization has yet unexamined impacts to our traditional biodefense frameworks.

On April 10 2019, the Center for the Study of Weapons of Mass Destruction (CSWMD) hosted a workshop to explore the growing digitization of biological data and its implications for biodefense. Held as part of CSWMD's project on Emergence and Convergence, this workshop supported ongoing work of the Interagency Working Group on Scientific Collections, at the request of the National Science and Technology Council of the Office of Scientific and Technology Policy within the White House. Three panels were convened to initiate discussion on the need for environmental baselines and standards, the challenge in using lists of "select agent" pathogens of concern, and how to achieve better global health security.

The workshop was enormously successful in providing a "first look" and assessment of how next generation sequencing and digitization is transforming our understanding of biodefense and biosecurity threats, and the role that scientific collections can play in identifying, understanding and mitigating these risks. Key recommendations stemming from the workshop emphasized standards, data quality, and the creation of 'functional taxonomies' in the interest of biodefense, particularly in light of advancing biotechnology capabilities. In addition, participants recommended the creation of collaborative resource sharing between collections and biodefense communities, as well as reinforced the importance of sample sharing in the interest of global health security.

Introduction

Biodefense and biosecurity in the United States rely on the ability to perform broad biosurveillance, to protect and secure biological agents of concern, as well as to diagnose and mitigate the potential consequences of the spread of global infectious diseases. Each of these abilities depends upon the collection and identification of biospecimens. As genomic sequencing and synthesis tools continue to grow, the genomic information associated with biospecimens is expanding rapidly. A scientific "renaissance" of Natural history is occurring through the increased capabilities of genomics, causing changes in our conception of species, further complicating differences between harmless, and what are currently considered biothreat organisms. The rapid convergence of biology with information technology has as yet unexamined impacts to our traditional operational biodefense frameworks.

On 10 April 2019, the Center for the Study of Weapons of Mass Destruction (CSWMD) hosted a workshop to explore the growing body of data associated with biological specimens and its implications for biodefense. Held as part of CSWMD's project on Emergence and Convergence[1], this workshop also supported ongoing work of the Interagency Working Group on Scientific Collections[2] (IWGSC) at the request of the National Science and Technology Council of the Office of Scientific and Technology Policy within the White House. Since 2006, the IWGSC has worked to assess the status and needs of the scientific collections owned, managed, and/or supported by the U.S. Federal Government, and to recommend ways to improve their management, effectiveness and impact. The goal for participants was to define questions, concerns, and potential opportunities to address between physical specimens and collections, genomic data, and biodefense tools and applications.

The event was attended by approximately 70 policymakers, analysts, and technical experts from across the U.S. government and other organizations. Importantly, it joined two sets of disparate stakeholders: those with expertise in systematics, taxonomy, and the curation and collection of biospecimens (typically with the goal of understanding biodiversity and ecology), with those from traditional security and defense fields who must detect and respond to bioevents. Three panels were convened to initiate discussion over a range of questions that were circulated to participants in advance. These included the need for baselines and standards, the increasing challenges in using lists of "select agent" pathogens of concern,[3] and finally, how specimens contribute to global health security. The two-way exchange between these separate communities revealed numerous opportunities for leveraging benefits on both sides. The participants also offered valuable suggestions for policy direction, surveillance technologies, and database options that could be pursued to biodefense advantage.

[1] https://wmdcenter.ndu.edu/Media/News/Category/13529/emerging-technologies/
[2] https://usfsc.nal.usda.gov/
[3] https://www.selectagents.gov/SelectAgentsandToxinsList.html

Baselines and Standards

Specimen collections and their associated genomic and phenotypic data are critical to establishing environmental baselines. The US must regularly monitor the environment for naturally occurring outbreaks, invasive species, as well as the purposeful malicious use of pathogenic or unsafe bioengineered organisms. Further, the US and others are interested in potentially beneficial synthetic biology applications, including gene drives to control infectious disease vectors, engineered microbes for bioremediation, biosensing and engineered viruses for wastewater treatment, to list a few. The US and others will need to surveil for engineered organisms against normal baselines, and to monitor the fate of engineered organisms in the environment, including their persistence and potential for gene transfer to wild organisms. All these areas represent biorisk in terms of misuse or unintended consequences.

A number of existing database resources were discussed, and participants noted the lack of standards for genomic data and metadata for specimens. Many public databases, amassed over decades when sequence tools were still evolving, contain data that is uneven in quality and can feature errors. The National Science Foundation (NSF) has spent the past decade on the digitization of its extensive biodiversity collections[4] – with the intent to extend the research value of such collections beyond the physical objects. They are focused on a model of "intelligent openness", i.e. resources should be accessible, assessable, intelligible, and usable. This is part of a larger theme, which is the need to integrate or link disparate databases into a usable set.[5]

A robust standards infrastructure is critically needed to ensure the quality of biospecimens, including standards for processing and banking biospecimens, measurement standards to evaluate quality, and data standards to enable interoperability and integration. High quality data and standards would accelerate the development and commercialization of reliable biotechnology products. In the absence of consistent quality and data, synthetic biology companies are assembling their own, often proprietary, databases – it was noted that incentives are needed for companies to accept consensus standards that could benefit the broader industry and community.

[4] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503559
[5] https://www.nature.com/articles/s41587-019-0080-8

Numerous international standards are under development by various Standards Development Organizations (SDOs), such as ISO and ASTM, as well as industry consortia and professional societies. ISO/TC 276: *Biotechnology* is one major global effort to develop standards for emerging biotechnology, including a comprehensive suite of standards for biobanking and biorepository, measurement standards, bioprocessing standards, and data standards. An example of a consortium effort to develop standards is the NIST Genome Editing Consortium,[6] which is working to develop measurement standards to quantify on- and off-target edits, data and metadata standards, as well as a unified lexicon.

A lack of standards and access to high-quality databases also creates problems for the development of screening tools and applications. Wild type baselines for both pathogen genotypes and phenotypes are crucial for the development of algorithms that can detect engineered strains, such as those being explored in the Functional Genomic and Computational Assessment of Threats (Fun GCAT) [7] program. If the wild type baseline data are inaccurate or incomplete, the machine learning training employed will 'learn' incorrectly – resulting in less useful biodefense tools. Interpreted broadly, poor standards mean unreliable field tests for military use.

[6] https://www.nist.gov/programs-projects/nist-genome-editing-consortium
[7] https://www.iarpa.gov/index.php/research-programs/fun-gcat

Challenges in Identifying "organisms of concern"

The tools and applications used to respond to complex biological events,[8] from initial detection through mitigation, rely primarily on lists of infectious organisms (for example, the Australia Group's Lists of pathogens or the CDC's Select Agents List). Today, as more genomic data is amassed for organisms, it is challenging traditional taxonomies and phylogenies. As noted above, in addition to challenges with slight changes in wild type strains that can occur naturally, great widespread sequencing of organisms is also leading to "identity" discrepancies for species and how organisms are genetically related. Systematic and taxonomic studies have shown that there can be minor differences in what might be considered a "threat' organism versus those which are not considered harmful. Further, synthetic biology can create chimeric and modified agents which do not fit easily into such lists. Scientific collections will be increasingly important as baseline vouchers and comparators to identify organisms with altered genetic sequences.

During this session panelists discussed the challenge of understanding the role of taxonomy in light of genomic data – and noted that genomics has been applied unevenly across species identification. Moreover, microbial systems are extremely genetically complex. The panelists used *Bacillus anthracis* as an example organism that demonstrates some of the challenges in taxonomy and the identification of what constitutes an "organism of concern". *B.* anthracis is part of a tight taxonomic cluster (the *Bacillus cereus* cluster) containing many species which are widely used in industry or which are ubiquitously present in the environment, but with only a specific subset having bioterrorism implications of use by state and non-state actors. In determining if a culture has *B. anthracis*, a chemical test can be used, but a single point mutation can change the results of that test, even though it is essentially still the same organism. This reveals the weaknesses of genotype-only classification; common tools for genotypic mapping (Digital DDH/Average Nucleotide Identity) will classify two organisms as same species if there is 70% match. However, many species of bacteria are extremely similar, making this threshold a potentially unreliable cutoff for distinguishing specific organisms of concern from harmless taxonomic neighbors. This of course can be further complicated by today's gene editing capabilities – exacerbating the question of "where to draw the line (of concern)" in the creation and implementation of future biodefense tools.

Scientific collections will continue to be needed in observing and cataloguing organisms, and to provide dynamic assessment of species longitudinally over time. Physical specimens represent a snapshot in time and location of a phenotype and genotype of an organism, but these aspects of organisms are not necessarily static and this has implications for digital knowledge. For example, changes to host range, abundance, and phenotypic variations may occur, as vectors and pathogens adapt to a changing climate. Or, wild-type and laboratory grown pathogens may show differences in the expression of proteins associated with their metabolic pathways,

[8] https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(17)30494-1/fulltext

complicating the problem of maintaining a physical reference sample that provides repeatable proteomic measurements. This begs the question of how similarly sequences should be to be classified the "same" species? Today's advanced sequencing tools could be generating an observer effect: how the sample is prepared presents a substantial risk of changing the sample itself.

The primary challenge is, "what blend of genotype/taxonomy/phenotype is best to use for which purposes?" Participants stressed that a "taxonomy of phenotypic expressions" is needed – focusing more on the functionality of pathogens of concern, instead of just their sequence identity – and these may vary for a purely scientific purpose, a surveillance purpose, or a regulatory purpose. Genotype and phenotype taken together is thus crucial to the discussion of what constitutes a pathogen. Infectivity, transmissibility, environmental stability, etc., are just some of the functions that make a microbe pathogenic. Pathogens vary in their relative pathogenicity, duration, onset, tropism, mortality rate, infectious dose, incubation period, and treatability – these levels of biological organization could be added to provide greater power to species' identification. Participants highlighted a set of important questions to focus on for future research studies in order to set up a genotype – phenotype taxonomy:

1. How do genetic differences manifest in phenotypic difference? What is the phenotypic diversity within each species?
2. Do we know the natural phenotypic diversity within a population of collected isolates, vs laboratory adapted/passaged isolates of a species?
3. Do we have a standard set of phenotypic assays that can be used to understand the phenotypic diversity of a species – including transcriptomic and proteomic characterization?
4. How do we integrate genomic data with metadata (phenotypic, proteomic, transcriptomic, etc.) to inform detection, diagnostic, medical countermeasure, and decontamination platform test and evaluation? How will this affect the effectiveness of platforms? Are we presently gathering such metadata in relevant scientific collections? How can the metadata be tracked beyond its initial access by the user?

Global Health Security

Scientific specimens are critical for global health security and response to outbreaks. Following on an earlier meeting sponsored by Scientific Collections International (SciColl)[9], the role of collections in mitigating infectious disease outbreaks has been described. A key aspect is zoonosis, i.e. many infectious outbreaks evolve from animal reservoirs of organisms that can spread to people. For example, DOD programs that engage in building health capacity in partner nations study tularemia, Ebola viruses, Nipah, meloidosis, paramyxovirus, etc., and their ecological cycling in host animals such as bats or macaques. The associated datasets for tracking infectious organisms around the globe continue to grow as well, as exemplified by such efforts as the Global Virome Project[10], and the Earth BioGenome Project[11].

Panelists reinforced the critical role of biospecimens and genetic sequence data (GSD) in identifying emerging infectious diseases (EIDs). This has recently been made even more important per the new National Biodefense Strategy[12], which includes naturally occurring outbreaks as well as perpetrated biological events, and has the specified goal to ensure biodefense enterprise preparedness to reduce the impacts of bioincidents. Additionally, the updated Global Health Security Agenda, aims to accelerate and optimize global health security[13]. Regarding pathogen sample sharing, anecdotally the lack of sharing of biological samples during outbreaks is often attributable to bureaucratic confusion rather than any intentional policy decision – and so one of the goals is to build awareness and capacities that emphasize speed, transparency, and systematic routines for international pathogen sharing.

Participants highlighted the "patchwork" of international agreements, frameworks, principles and policies that have developed over the past few decades. For example the WHO's Global Influenza Surveillance and Response System (GISRS)[14], facilitates sharing of influenza strains, beginning as far back as 1952 and fosters global confidence and trust in public health across the GISRS network. The more recent Pandemic Influenza Prevention (PIP) Framework calls for pandemic influenza strains and their genetic sequence data to be shared
Certain recent developments in other international fora could affect the ability of the international community to effectively prepare for and respond to outbreaks that threaten global health security.

[9] http://www.pnas.org/content/113/1/4
[10] https://www.globalviromeproject.org/
[11] https://www.earthbiogenome.org/
[12] https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Biodefense-Strategy.pdf
[13] https://www.ghsagenda.org/ghsa2024
[14] https://www.who.int/influenza/gisrs_laboratory/en/

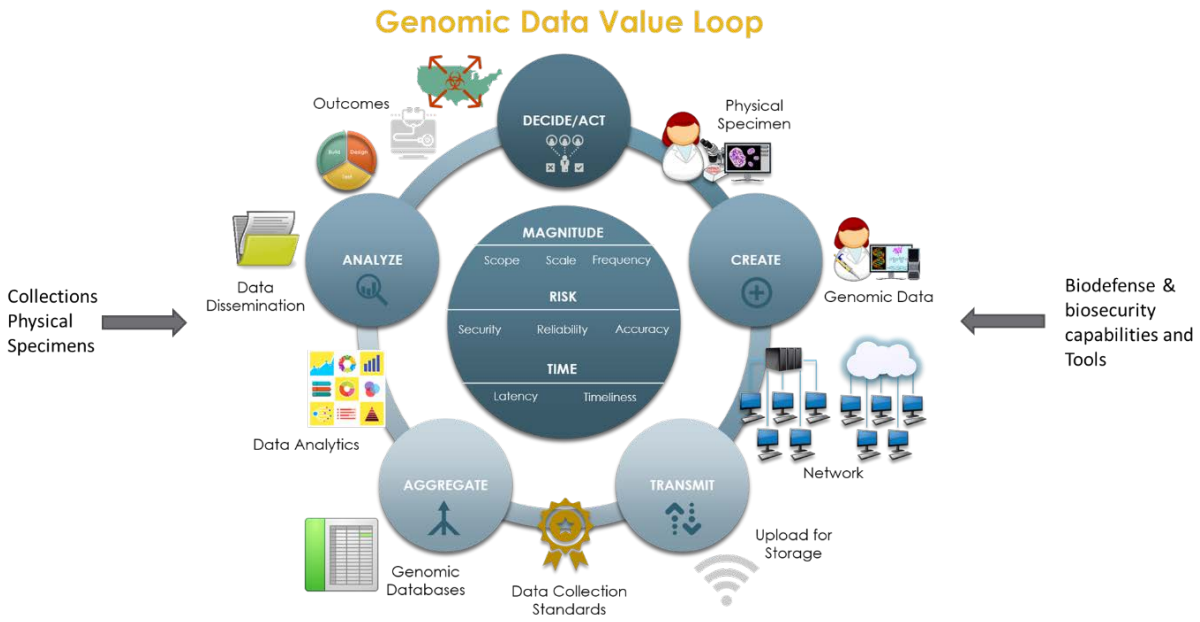
The Nagoya Protocol to the Convention on Biological Diversity (CBD), stipulates Parties should set clear conditions for access to genetic resources and clear conditions and provisions for what, if any, benefit sharing is required for benefits derived from the utilization of those resources. The Nagoya Protocol requires each Party to take measures to certify that any genetic resources utilized in its jurisdiction has been accessed in accordance with the other relevant Party's ABS measures. Some countries are choosing not to require benefit sharing in exchange for access to pathogens. Other countries are implementing domestic measures that regulate access to genetic resources, with some also choosing to implement restrictions on access to pathogens and genetic sequence data (GSD). For example, Brazil is using a broad interpretation of "genetic heritage" to capture GSD in its domestic ABS measures. Parties to the CBD and the Nagoya Protocol are still discussing whether and how GSD should be addressed in relation to those instruments, but regardless of that discussion, countries can choose to implement national ABS measures that cover GSD. The United States signed but has not ratified the CBD and is not a Party to it or its Nagoya Protocol.

In the public health sphere, the United States advocates for timely sharing of pathogen samples needed to prepare for and respond to bioevents, as well as the open sharing of genomic data and information in the interest of global health security. All panelists agreed that physical samples will be needed for some time, but that genome sequence data and the field use or point of care next generation sequencing tools represent the "future" of global health biosurveillance.

Summary and Recommendations

The workshop was enormously successful in providing a "first look" assessment of how digitization is transforming the role of scientific collections and their use in biodefense and biosecurity. Biological digitization, or the conjoining of biological physical and cyber realms, can best be represented by a "genomic value data loop" between specimens and biodefense operational tools. (See figure below). Each point in the data loop can represent points of need or where actions could be taken to better enable the use of biospecimen data in the interest of biodefense. Importantly, areas identified for data improvement necessarily track back to the how biospecimens should best be collected and maintained for their optimized use and benefits.

Recommendations stemming from the workshop revolved around several key areas. Within the "genomic data loop", participants emphasized standards, data quality, and the creation of 'functional taxonomies' in the interest of biodefense. In addition, participants recommended the creation of collaborative resource sharing between collections and biodefense communities, as well as reinforcing the importance of sample sharing in the interest of global health security.



Standards

Standards were recognized as a critical need for physical specimens, data collection and databases. Standards are an underlying need for the creation of genotype/phenotype taxonomies, the overall quality of data, and data, and resource sharing (see below).

Data Quality

There was consensus that existing pathogen genomic databases contain errors, inaccuracies, and incomplete information. This limits the capabilities of field detection, regardless of the tools employed, and further, if machine learning or complex computational tools are to be applied to detect novel or engineered biothreats, they must be predicated on accurate baseline data. A set of environmental standards (for both the collection of physical specimens as well as data processing and analytical standards) should be applied to specified, manually well-annotated databases. Participants were unable to agree as to whether existing data should be compiled, given known inaccuracies, or if it would be better to 'start from scratch', however the following recommendations were made:

- Explore incentives for the creation of dynamic, annotated datasets; explore incentives for companies to share their quality sequencing data.
- Focus on creating the best documented collections possible, whether the sample is physical or electronic.
- Explore computer testing in the virtual "build" of organisms, thereby saving the information, and creating more robust statistical databases;

Need for a "functional taxonomy" that provides for understanding how genotype relates to functional phenotype.

Genomic sequence data (GSD), while having utility in tools, must be supplemented by metadata (metadata can include phenotypic data, transcriptomic data, proteomic data, location of collection). Included in this should be understanding generalized functions for pathways that pathogens use, and the parallel systems that are enjoined in hosts. The government's ability to establish policy, standards, etc. around specimens and biodefense tools, might be better served by examining known pathways. Recommendations include:

- Deeper dives on questions articulated in session 2 (above)
- Engage with the biotechnology industry, which is already heavily invested in this topic; for example the cell and gene therapy industry could advance this capability.
- Determine if this can assist in the ongoing deliberations on DNA synthesis screening.

Collaborative dialogue and resource sharing is needed.

Participants expressed the desire to further the collaboration between the biodefense, the scientific collections, and the diplomatic communities on all recommendations made in the report. The biodefense community may build tools that could be shared back to the biodiversity and collections community; the collections community could provide insights that benefit tool development; the diplomatic community could share best practices from other fora with regards to international sharing of genetic sequence data and physical samples. Recommendations include:

- Better use of existing interagency processes, or creation of new interagency groups to cover these issues;
- Establishment of "communities of practice" for particular organisms – it was suggested that there may be precedent for this in other fields;
- Identification of international best practices for genetic sequence data and physical samples;
- It is clear that scientific biospecimen collections are intrinsic to important components of biodefense, as well as successful biotechnology; as such they contribute to these as critical infrastructures.
- Explore venues to demonstrate the value of collections to policy makers and funders;
- Create venues for data sharing, not only across the USG but that include international inputs (see above).

Global sharing of physical samples, along with genetic sequence data, is still a critical component of global health security, and frameworks for such sharing are important.

Although the topic has been noted in other forums, participants at this workshop re-emphasized the importance of both pathogen and possibly non-pathogen sample sharing as an integral critical component of global health security. Some specific recommendations that could be taken on by this particular community of interest might include:

- Exploration of collaborative mechanisms for physical and digital collections (in sharing and uses for health security and biodefense);
- International genetic sequence data sharing is currently not always reciprocal, so efforts should be made to expand existing efforts [15] and determine the adverse impacts of lack of sharing;
- Creation of resource sharing platforms across international boundaries - these could be modeled after existing platforms which have been successful for influenza[16],[17] and antibiotic resistant bacteria[18]

[15] Global Genomic Medicine Collaborative, https://g2mc.org/
[16] Centers of Excellence for Influenza Research and Surveillance (CEIRS), http://www.niaidceirs.org/resources/data-sharing/; includes Influenza research database: https://www.fludb.org/brc/home.spg?decorator=influenza
[17] Global Initiative on Sharing All Influenza Data, https://www.gisaid.org/ .
[18] Pew's engine SPARK - Shared Platform for Antibiotic Research and Knowledge. https://www.pewtrusts.org/en/about/news-room/press-releases-and-statements/2018/10/22/achaogen-provides-data-to-spark-pews-platform-for-antibiotic-discovery-research

Relevant Literature:

Baker, K. et al. Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species. *Nature Communications*, 2018; 9 (1) DOI: 10.1038/s41467-018-03949-8. https://www.nature.com/articles/s41467-018-03949-8

Boundy-Mills, K., et al. The United States Culture Collection Network (USCCN): Enhancing Microbial Genomics Research through Living Microbe Culture Collections. *Appl. Environ. Microbiol*. Aug 2015, 81 (17) 5671-5674; DOI: 10.1128/AEM.01176-15. https://aem.asm.org/content/81/17/5671.short

DiEuliis, D., Johnson, K., Morse, S., Schindel, D. Opinion: Specimen collections should have a much bigger role in infectious disease research and response. *PNAS* January 5, 2016 113 (1) 4-7; https://doi.org/10.1073/pnas.1522680112. Or http://www.pnas.org/content/113/1/4

Garrity, G. M. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? *Journal of Clinical Microbiology* Jul 2016, 54 (8) 1956-1963; DOI: 10.1128/JCM.00200-16. https://jcm.asm.org/content/54/8/1956

Hamilton, H, Guala, G., and Simpson A. Promoting synergy in the innovative use of environmental data—Workshop summary. (2018) https://pubs.er.usgs.gov/publication/ofr20181104

Kuhn, et al. Classify viruses, the gain is worth the pain. *Nature*, vol 566. February 21, 2019. https://www.nature.com/magazine-assets/d41586-019-00599-8/d41586-019-00599-8.pdf

Lewin, H. A. et al. Earth BioGenome Project: Sequencing life for the future of life. *PNAS*, Apr 2018, 115 (17) 4325-4333; DOI: 10.1073/pnas.1720115115. https://www.pnas.org/content/115/17/4325.long

Schindel, D.E., and Miller, S. E. Provisional Nomenclature The On-Ramp to Taxonomic Names. In Systema Naturae 250: The Linnaean Ark; Chapter 10, 2010. https://www.researchgate.net/publication/279400233_Provisional_Nomenclature

Nilsson, R. et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, 2018. doi: 10.1093/nar/gky1022. https://academic.oup.com/nar/article/47/D1/D259/5146189

Reardon, S. How machine learning could keep dangerous DNA out of terrorists' hands. *Nature*, 31 January 2019. https://www.nature.com/articles/d41586-019-00277-9?utm_source=Nature+Briefing&utm_campaign=79dd080757-briefing-wk-20190201&utm_medium=email&utm_term=0_c9dfd39373-79dd080757-42718491

Siddell, Stuart. Why virus taxonomy is important. Microbiology Society, February 13, 2018. https://microbiologysociety.org/publication/past-issues/imaging/article/why-virus-taxonomy-is-important.html

Stackenbrandt, E. Diversification and Focusing: strategies of microbial culture collections. *Trends in Microbiology*, Volume 18, Issue 7, July 2010, Pages 283-287. https://www.sciencedirect.com/science/article/pii/S0966842X10000764

Venkateswaran, K., et al. Non-Toxin-Producing Bacillus cereus Strains Belonging to the B. anthracis Clade Isolated from the International Space Station. *mSystems* Jun 2017, 2 (3) e00021-17; DOI: 10.1128/mSystems.00021-17.

Zhang, S. New DNA Database Allows Far Faster Searches for Pathogen Genomes. *Defense One*, February 4, 2019. https://www.defenseone.com/technology/2019/02/new-dna-database-allows-far-faster-searches-pathogen-genomes/154633/?oref=defenseone_today_nl

Resources on the Web:

- Interagency Working Group on Scientific Collections and clearinghouse: https://usfsc.nal.usda.gov/
- National Ecological Observation Network (NEON), supported by NSF through Battelle: https://www.neonscience.org/
- Biodiversity Information Serving Our Nation (BISON): https://bison.usgs.gov/#home
- IDigBio: https://www.idigbio.org/
- Dept. Of Energy Joint Genome Institute (JGI) https://jgi.doe.gov/
- NIST Biosystems and Biomaterials Division: https://www.nist.gov/mml/bbd
- Finding Engineering-Linked Indicators https://www.iarpa.gov/index.php/research-programs/felix
- American Type Culture Collection: https://www.atcc.org/
- National Biodefense Analysis And Countermeasures Center (NBACC) fact sheet: https://www.dhs.gov/publication/national-biodefense-analysis-and-countermeasures-center-nbacc
- Nagoya Protocol on Access and Benefit Sharing https://www.cbd.int/abs/
- Pandemic Influenza Preparedness (PIP) Framework https://www.who.int/influenza/pip/en/
- Global Virome Project: https://www.globalviromeproject.org/
- Earth BioGenome Project: https://www.earthbiogenome.org/