Global commercialization trends of microbial products and processes

Charles Parker[1], Grace Rodriguez[1], David P. Labeda[2] and George M. Garrity[1,3]
[1]NamesforLife, LLC, East Lansing, MI, [2]USDA Agricultural Research Service, Peoria, IL,
[3]Michigan State University, East Lansing, MI

NamesforLife
Bringing meaning to life...

Project Goals

A fundamental premise in industrial microbiology and biotechnology is that microorganisms are a source of commercially useful products and processes. While there is always an element of serendipity in the discovery process, success is often defined by access to appropriate resources, including collections of well-documented strains and historical knowledge about their metabolic and genetic potential. While some organizations still maintain private strain collections many have abandoned this strategy due to cost. There are, however, public resources that can serve as a substitute, chief of which are the international patent repositories. These collections hold the strains that are key to many patented microbiological inventions. Once a patent expires, these strains can be released without restriction. These strains are distinct from non-patent strains in public collections because of the amount of information that is publicly available, although that information is difficult to obtain or use. Our objective is to make the connections between strains and the patent literature easy to navigate and to make the information about patented microbial products and processes more readily discoverable. We recently completed a first pass through the USDA ARS Patent Collection (NRRL Collection, Peoria, IL). Using proprietary text mining methods, we were able to identify global commercialization trends in 162 technology classes over a 70 year time span by following more than 4,000 distinct NRRL strains referenced by over 16,000 US and foreign patents drawn from a corpus of >80 M patent documents.

Background

Patent repositories

Microorganisms and their metabolic products and processes are protectable intellectual property in most countries. Securing a patent on a microbial invention requires filing an application in each jurisdiction where protection is sought in which the inventor(s) must establish the novelty, utility and non-obviousness of the claimed invention. Their detailed description must fully disclose the nature and workings of the invention so that it enables others who are "skilled in the art" to replicate the invention. Failure of enablement is grounds for revocation of the rights granted under a patent.

Microbiological inventions are distinct from chemical, mechanical and electrical inventions in that enablement of the invention typically requires the live organism to fulfill the legal requirements of full and complete disclosure. Since 1949, the USPTO has required that viable samples of the subject microorganism be deposited in a public culture collection in conjunction with the filing of a patent application. Similar rules went into effect in most other industrialized nations soon after, leading to the establishment of a number of national repositories that held collections of these preserved microorganism that would become available when an issued patent was presented. In 1980, the process of depositing microorganisms in association with patent filings was simplified with the ratification of the Budapest Treaty, which allowed a single deposit in one of 12 international repositories (now 37) to suffice for all patent applications made under the Patent Cooperation Treaty. The USDA ARS Patent Culture Collection in Peoria, IL is one of the oldest patent repositories and was among the original 12 international repositories.

The patent literature

The patent literature represents a distinct and useful body of scientific and engineering knowledge that exists in parallel to the scientific, technical and medical (STM) literature that is familiar to most researchers. Both bodies of literature are intended to provide an accurate source of factual information that documents the priority of discoveries and establishes the links to prior knowledge. However, the manner in which these facts are reviewed and verified differ significantly. Whereas the STM literature is designed to meet the needs of a given community of practice and can vary widely in scope and form, the patent literature is designed to fulfill specific legal needs and, when granted, provides a bundle of legal rights to the owners or assignees of the patent. Both the STM and patent literature undergo a careful review, but the review processes differ dramatically. Both are indexed and catalogued extensively so as to make documentation of prior work more readily discoverable, but the manner and extent to which the indexing and cataloging is done also differs. Both bodies of literature also have associated metadata, which aids in the indexing and cataloging process as well as in retrieval of subsets of related documents, but here too, the differences are significant. Whereas indexing and cataloguing of the STM literature is an activity of a mixture of players in the public and private sector applying a variety of methods to different portions of the literature, indexing and cataloging of the patent literature is driven largely by national and international intellectual property offices that apply standardized methods to major blocks of the patent literature. These rich metadata become a part of the public record and are integrated directly into the patent documents.

When available in digital form, both the STM and patent literature can also be operated on with a variety of text-mining tools, such as those developed by NamesforLife. In particular, these tools overcome the limitations of conventional indexing and annotation methods that typically use controlled vocabularies or key-words that are applied by human indexers to documents as they are being read and classified. The term and concepts applied by indexers are generally not used by those who write, read or search for documents, thus making such systems difficult to use and subject to anomalous results. Our approach makes use of externally managed terminologies and nomenclatures (subject language terminologies; SLT) that are commonly used by practitioners and have precise and known meaning at a given point in time. Our tools allow us to index and classify documents, to extract this specific type of contextual information from any collection of documents, and to use this informaition to augment prior classifications. This can result in significant improvements in the recovery of related documents. It also provides a method to facilitate connections between the literature and records contained in databases, such as those maintained by the ARS Patent Culture Collection; forming links between a deposited culture and what is known or might be inferred about that organism based on the patent literature.

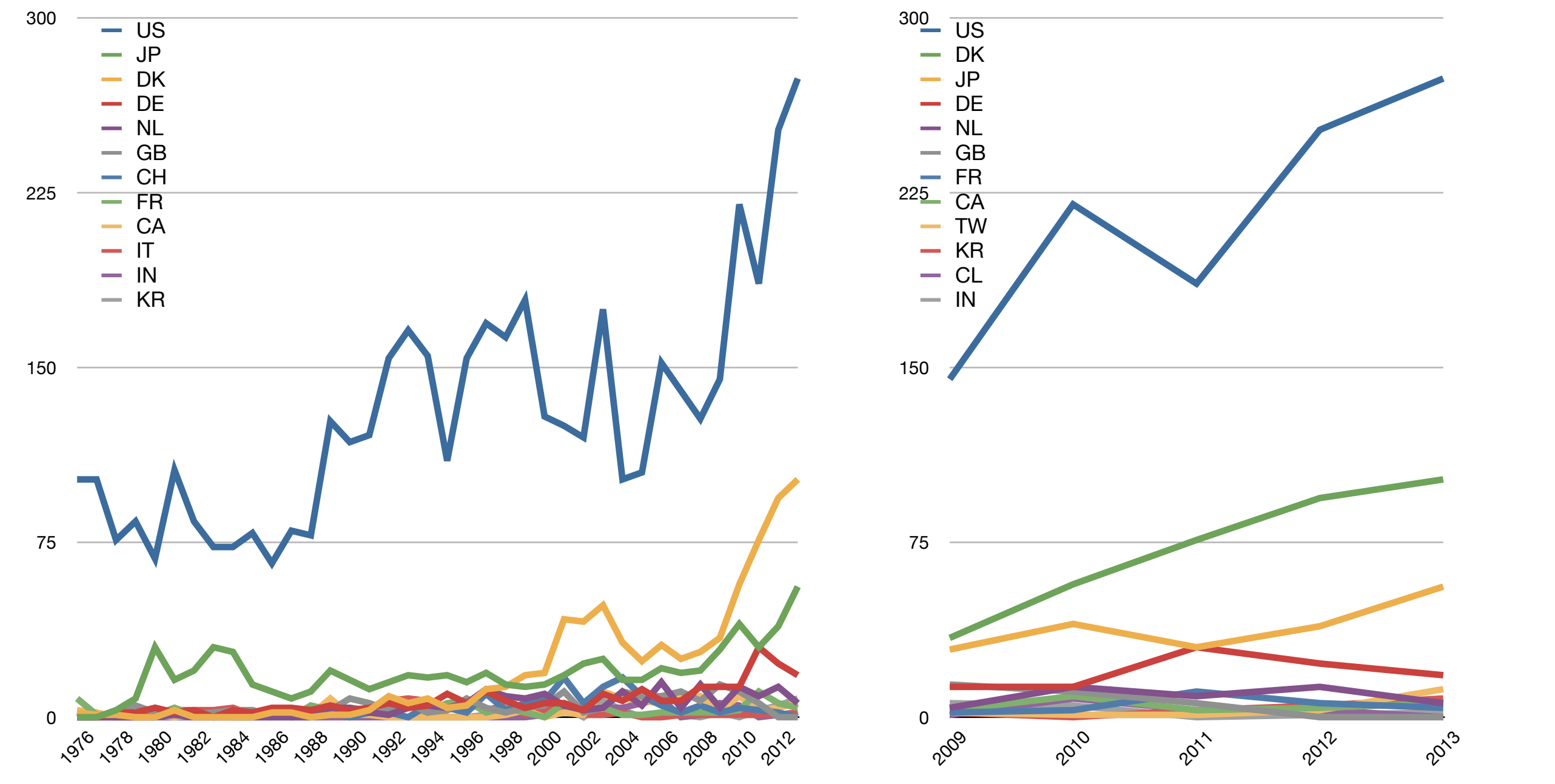
The goal of this project was to establish which cultures in the holdings of the USDA ARS Patent Culture Collection have been referenced in one or more US or foreign patents and to perform a meta-analysis of the resulting corpus in which the recovered patents and applications would be grouped based on the subject organisms (at the genus level), technology groupings and time of publication as a prerequisite to text-mining and development topic related SLTs.

Metadata extraction

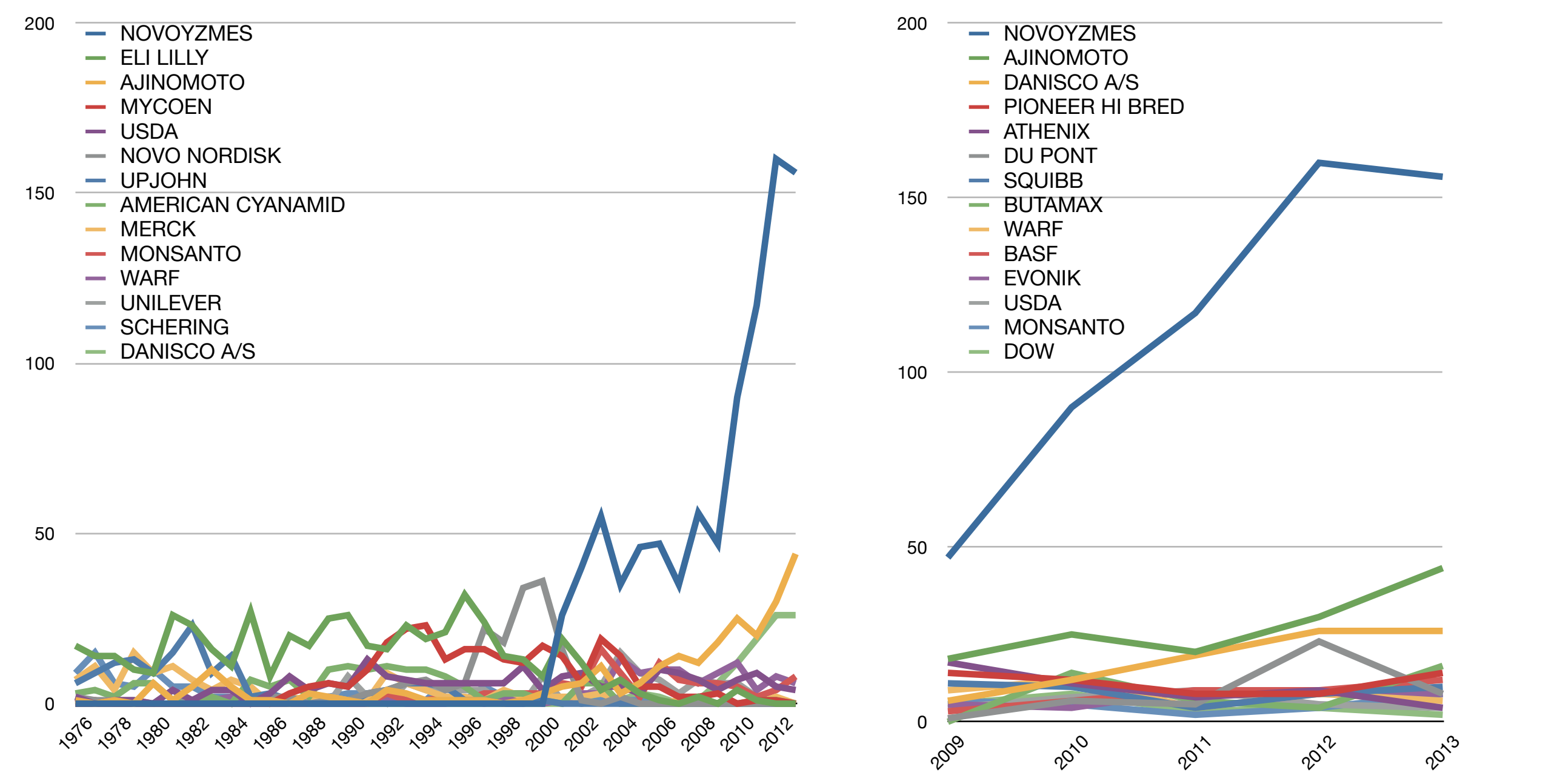
The NRRL corpus was created by extracting a subset of patents from the Alexandria Database (IFI Claims/Fairview), which consists of approximately 80M patent grants and applications from 72 patent authorities. The initial corpus consisted of 16,088 documents from eight national/international patent offices (Austria, Belgium, Switzerland, China, European Patent Office, Spain, US and the World Intellectual Property Office). The initial survey query was based on various permutations of the the collection accession numbering schemes. The resulting query was used to search Alexandria (1976 - 2013) and Google Patents (1900-1975). Redundant UCIDs were filtered from the list and the resulting full documents were retreived from Alexandria (XML) or Google (PDF->OCR (text)->XML). The resulting content was then normalized, and parsed to extract the patent metadata (e.g., UCID, Assignee, IPC classification codes, country codes, publication date). Strain identifiers were extracted, manually normalized and checked for equivalencies. These data were merged with the NamesforLife taxonomy database to determine the correct nomenclature.

Meta-analysis

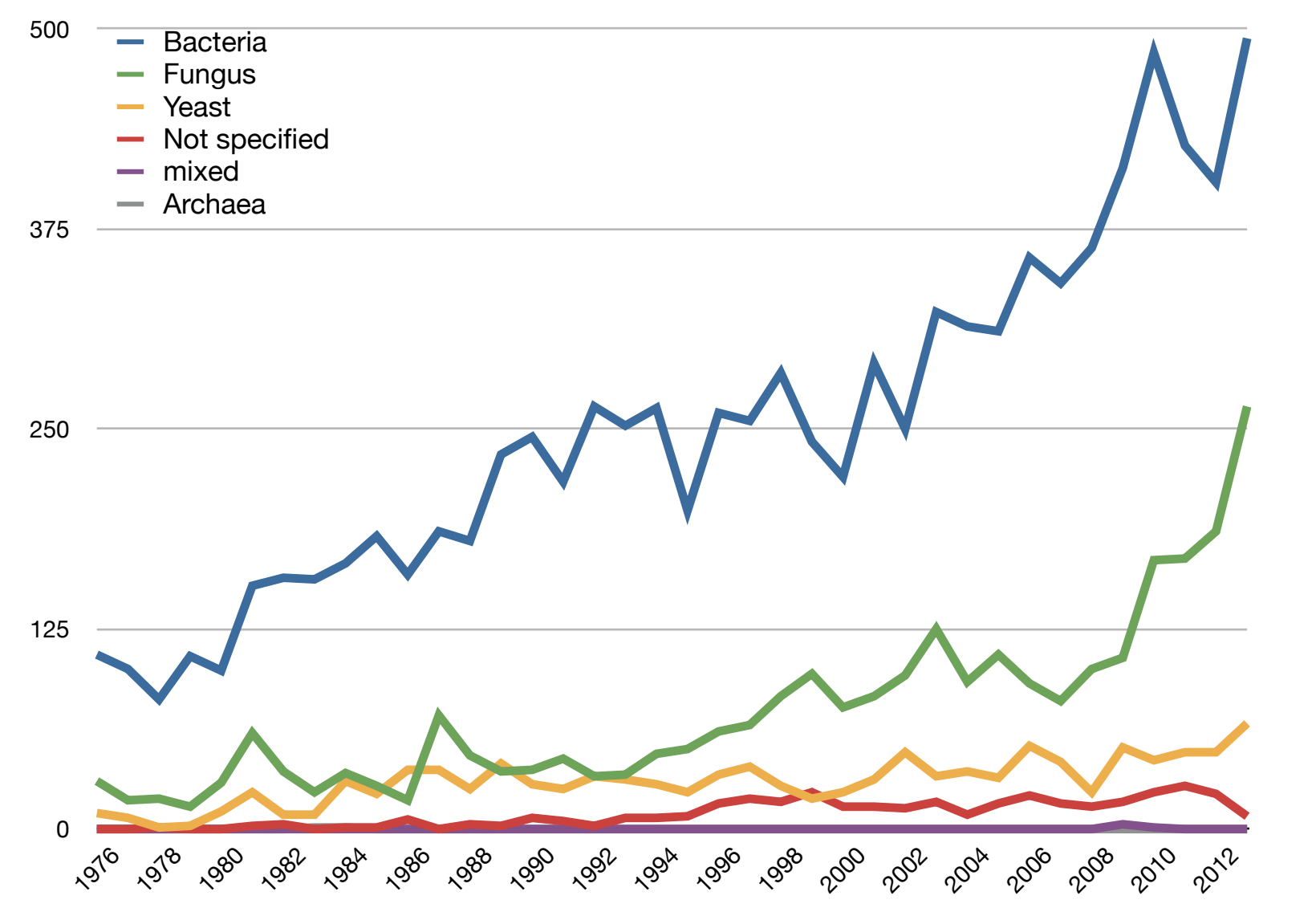
The extracted metadata was loaded into the R statistical computing environment to explore some of the general trends in commercialization of microbial products and processes associated with the deposits in the USDA ARS Patent Culture Collection. Semiotic fingerprint analysis was also performed in which the subset of patents for which both strain identity and IPC codes were available (n=10,784). This resulted in the placement of the patents into 158 clusters, ranging in size from 2 - 865 documents.



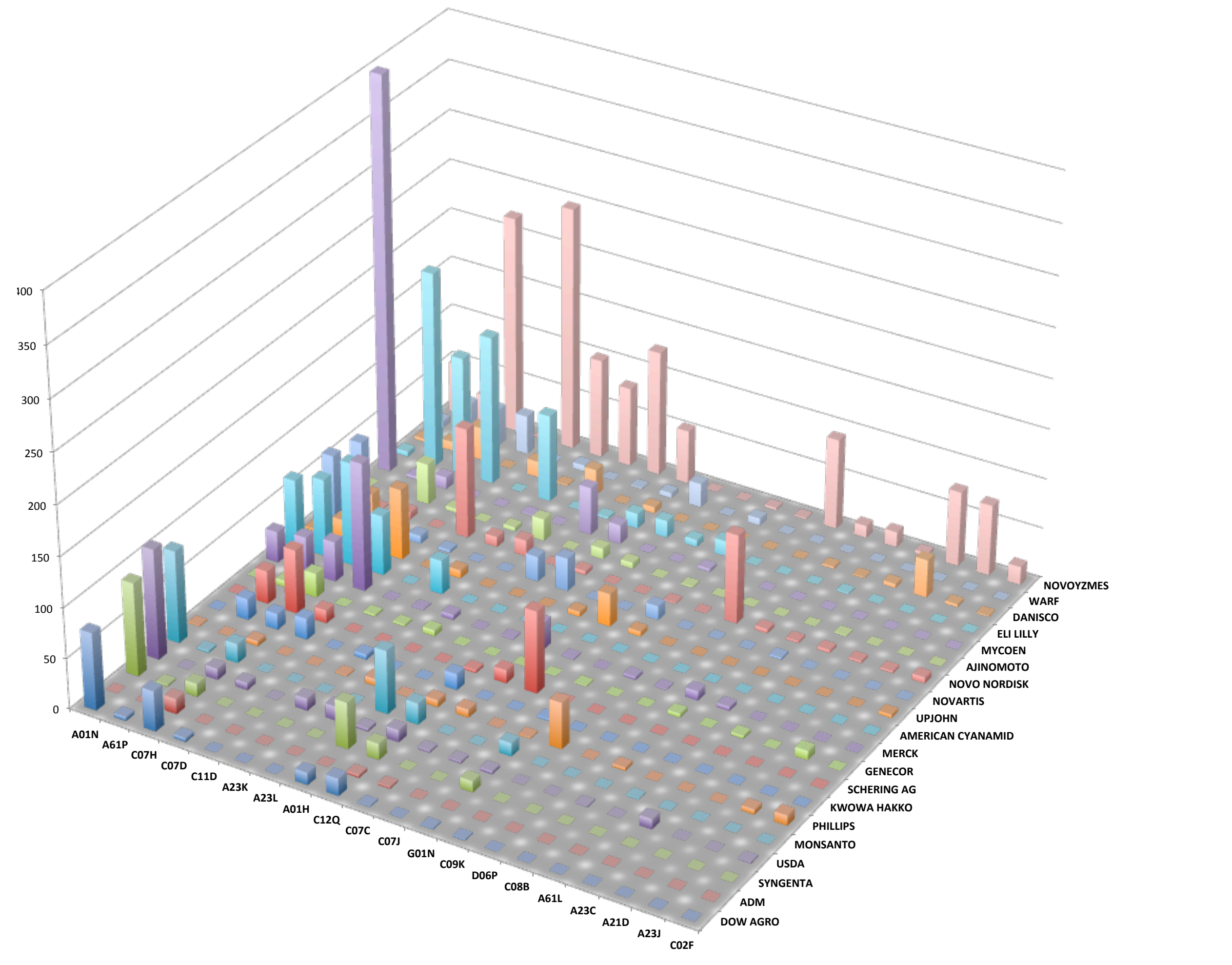
Origin of the NRRL patent strains by country. The top 10 countries represent >95% of the total number of identifiable patent strains for which these data were recorded in the patent metadata (n=7944) from 1976-2013. The right panel shows the current trend. The total number of countries cited in the patent record during that time period was 51 (1976 - 2013) and 38 (2009 - 2013). The number cited deposits has increased steadily from 142 to 518 annually.



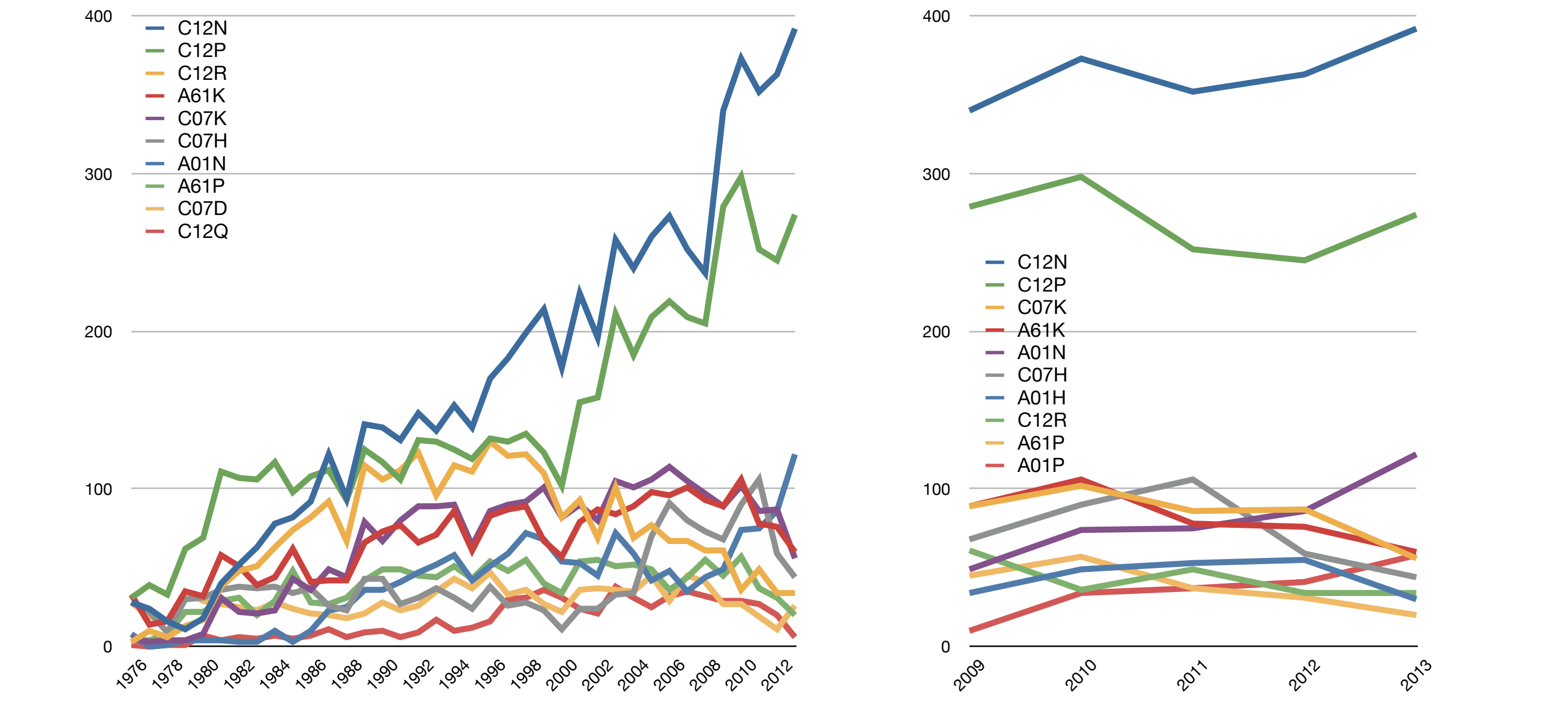
Origin of the NRRL patent strains by assignee. The top 10 assignees account for >40.6% of the total identifiable assignees in the patent record. Left panel, 37 year trend; right panel, five year trend. The total number of assignees cited in the patent record during that time period was 912 (1976 - 2013) and 324 (2009 - 2013).



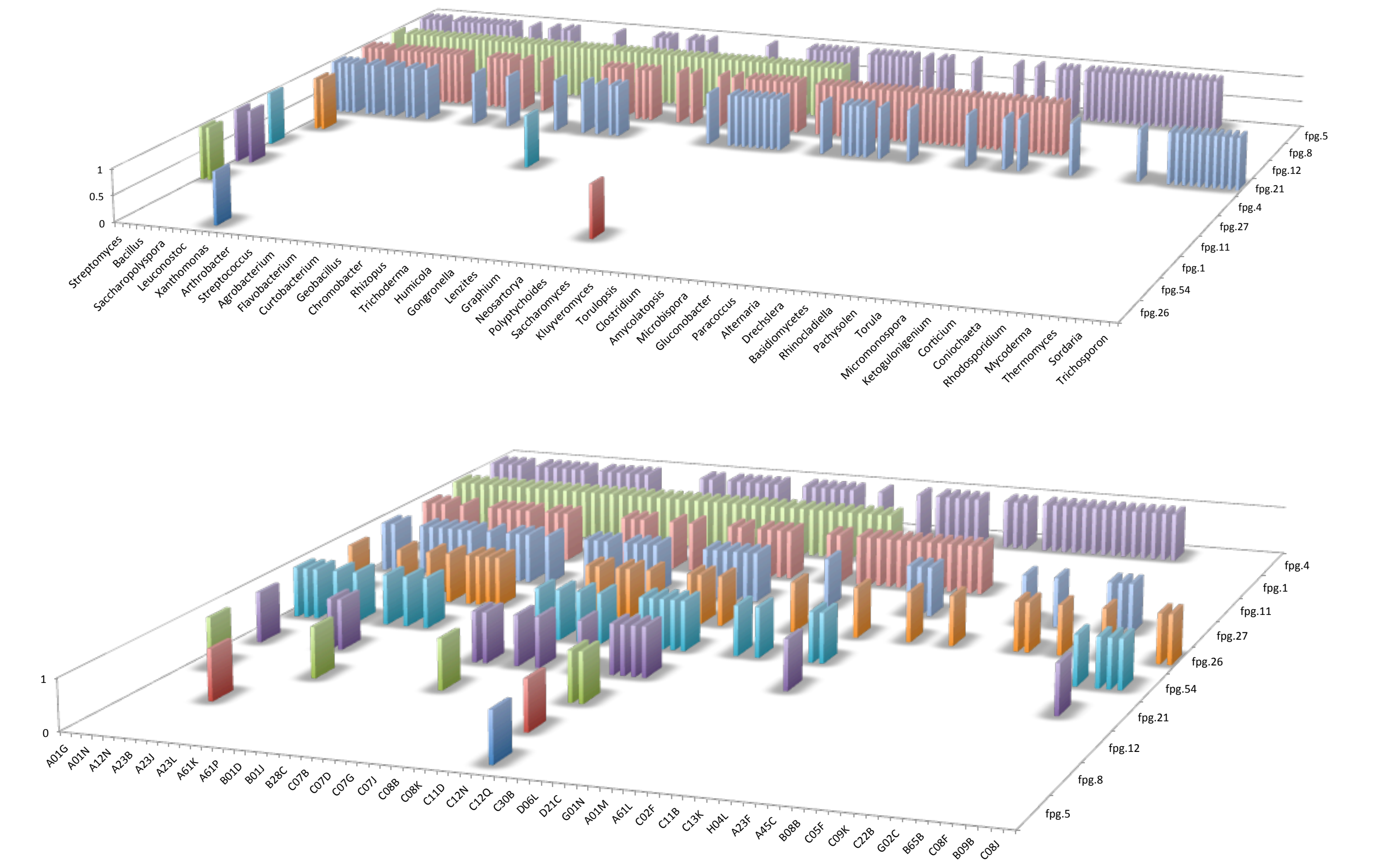
Taxonomic identity of the NRRL patent strains. Bacteria and filamentous fungi have represented the dominant taxonomic grouping of NRRL patent strains throughout the history of the collection, despite major shifts in technology. They account for >89% of the referenced deposits. The predominant genera of bacteria cited are *Escherichia*, *Bacillus*, *Streptomyces*, *Corynebacterium* and *Saccharopolyspora*. *The predominant* genera of filamentous fungi are *Thielavia*, *Aspergillus*, *Trichoderma* and *Penicillium* and *Fusarium*. The total number of all cited genera is 284.



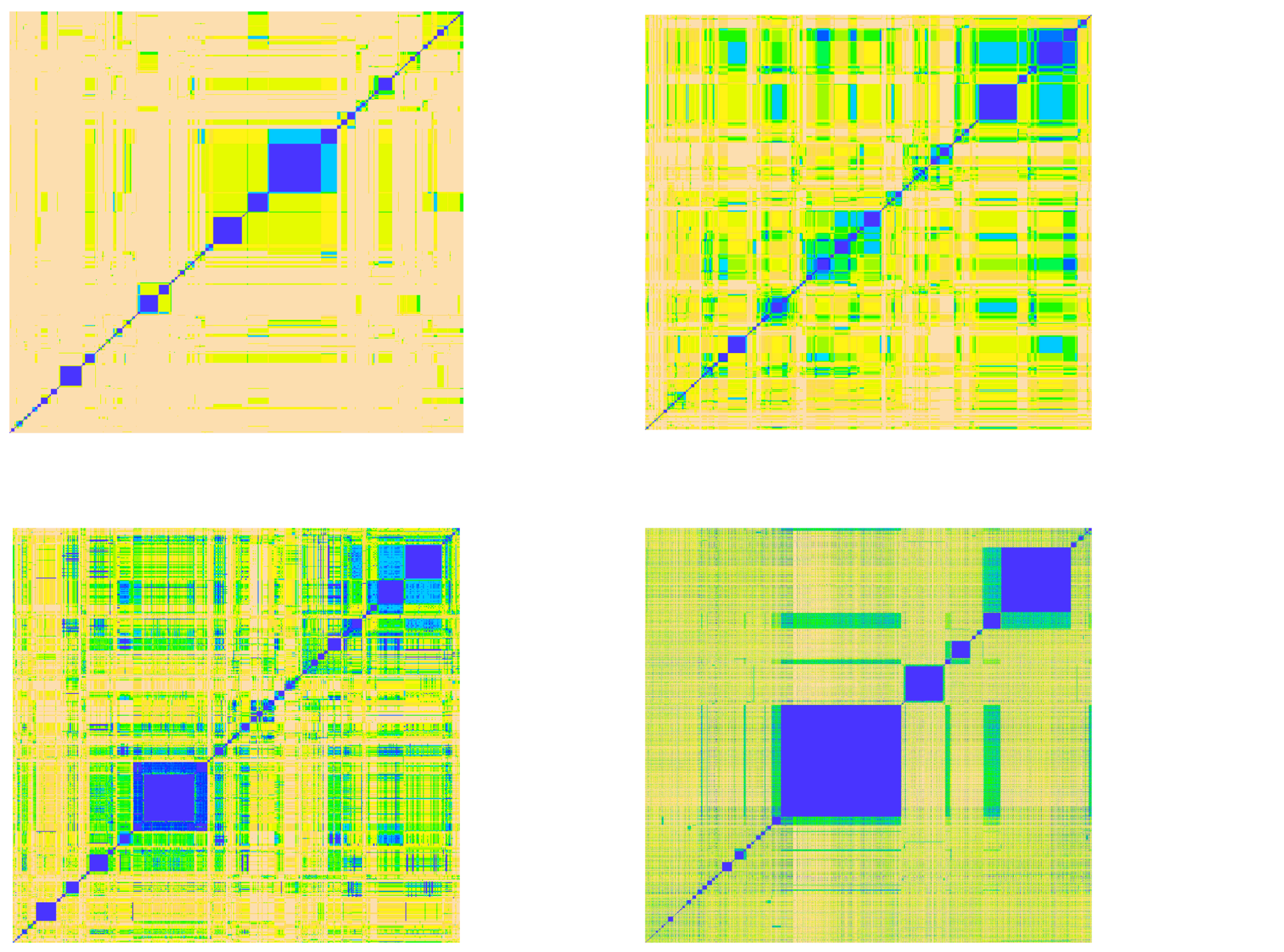
Association of NRRL patent strains by assignee and IPC Code. Patent metadata is a useful source of business intelligence as well as technical knowledge. When patent classification codes are combined with assignee data and other information that can be extracted from patents and external sources, it is possible to infer a great deal about the research and commercialization activities of a given organization. Here, we show the top 20 IPC classification codes associated with referenced patent strains for the top 20 assignees. Note the strong grouping among different industry representatives.



Association of NRRL strains with various IPC technology classes. During examination each patent is assigned to one or more different technology classes. When combined with information about identity of the patent strains, this provides useful information about metabolic capabilities of each organism that can be used for a variety of purposes. Trend analysis of various technology classes can also provide some indication as to which technologies are maturing (e.g. C12R, A61K) and which are emerging (A01N, A01P).



Semiotic fingerprinting the NRRL patents. Most patents are placed into multiple technology classifications during prosecution. Likewise, many different taxa may be associated with the same technology (e.g., production of antimicrobials or bioethanol). Both may also be affected by terminologies that are fluid and laden with synonyms an polysemes. This significantly confounds search and retrieval. To address this problem, we have developed a novel method of indexing and filtering text that leverages this complexity to rapidly cluster large collections of documents and to rapidly organize them into contextually relevant categories that can be further operated on by machines or presented



Visualization of the NRRL patents. The metadata extracted from the subset of 10,784 patents that could be positively associated with NRRL patent strains and for which IPC classification codes were also available were subjected to a series of sequential cluster analyses. The results were then projected as heatmaps of the input similarity matrices that were re-ordered along both dimensions according to the output of the clustering. This reveals the topology of the data and serves as a means of indexing the related patents. The similarity matrices are symmetric, with identity (pairwise match to self) along the diagonal. Highly similar results are colored dark blue and should be found along the diagonal. Dissimilar pairings are tan-colored. Intermediate levels of similarity follow a gradient from blue -> green -> yellow -> tan. Top left, classification based on taxonomic identity of patent strain(s); top right, classification based on IPC codes of patents; bottom left, classification based on the taxonomic identity × IPC codes (a semiotic fingerprint); lower right, a semiotic fingerprint of the 7,994 patents for which assignee metadata was available.

Future Work

With this retrospective meta-analysis now completed, we are exploring the utility of automatically generating periodic updates to clients about emerging technology trends and the availability of strains following patent expiration or abandonment. We are also exploring the possibility of extending this approach to other international patent repositories as part of an alerting service. This meta-analysis also serves as a preliminary step in our terminology and ontology development pipeline.

Acknowledgments

Funding for this project was provided through the DOE SBIR/STTR program (DE-SC0006191). Funding for the NamesforLife infrastructure was received from the DOE SBIR/STTR program (DE-FG02-07ER86321), the Michigan Small Business Technology Development Corporation, the Michigan Strategic Fund, the Michigan Economic Development Corporation, and the Michigan Universities Commercialization Initiative.