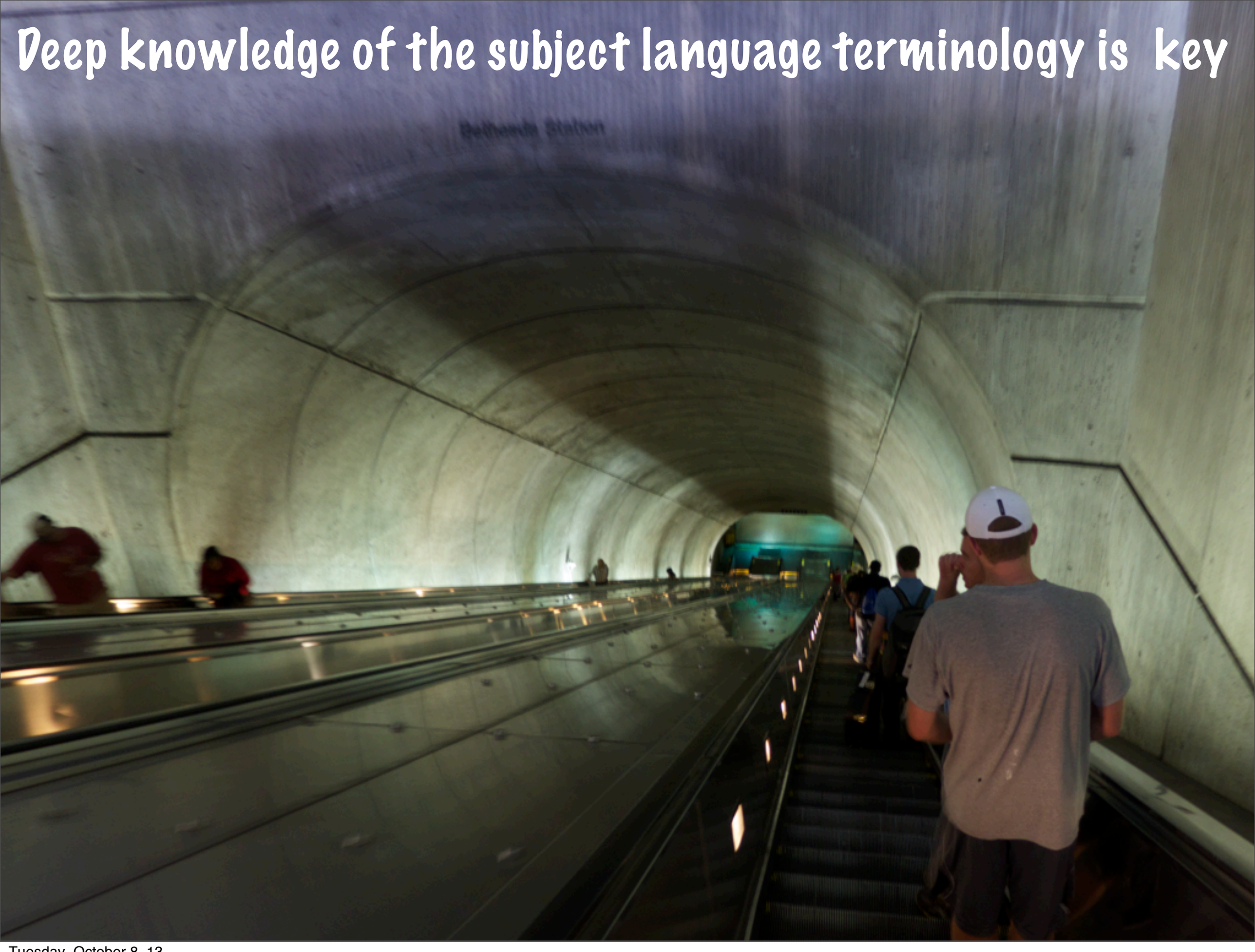
Text-mining, metegenomics, names and phenotype: A high-level view.

George Garrity
Michigan State University
NamesforLife, LLC


Zipf's law applies
...but so do the rules of nomenclature
...and communities of practice
...and the principles of semiotics and taxonomy

Deep knowledge of the subject language terminology is key



"Kbase also should incorporate metabolic, physiological, and morphological phenotypes used to identify species (e.g., from Bergey's "differential characteristics" tables)."

From the DOE Systems Biology Knowledgebase Implementation Plan

Description of *Legionella steelei* sp. nov.

Legionella steelei (steel'e.i. N.L. masc. gen. n. *steelei* of Steele, pertaining to the Australian microbiologist Trevor Steele, who performed pioneering work on the ecology and pathogenesis of *L. longbeachae* infection).

Gram-negative-staining rod. Grows on BCYEα agar, but not on tryptic soy blood agar or BCYEα agar without L-cysteine. Optimal growth temperature is 35 °C, with no or very poor growth at ≥37 °C. Produces large amounts of diffusible yellow-green fluorescent pigment. Variable production of colony-bound blue-white fluorescence. Positive in tests for activities oxidase, catalase, gelatinase, hippurate hydrolysis and β-lactamase. Negative result for glucose utilization. Motile, with monopolar flagella. Dominant cellular fatty acids are anteiso15:0, 16:1 and 16:0. Virulent in *Acanthamoeba castellani*. 16S rRNA, *mip*, *rpoB*, *rnpB* and *proA* gene sequences differ significantly from all recognized species of the genus *Legionella*.

The type strain, IMVS-3376^T (=IMVS 3113^T=ATCC BAA-2169^T), was isolated from a human respiratory tract specimen.

Major features included in the NamesforLife Phenotypic Index

Strain metadata
- N4L Exemplar ID
- Isolation source
- Isolation method
- Isolation substrate
- Geographic location
- Environmental information
- Host
- Strain designation
- Collection ID(s)
- Taxon status (type/non-type)

Genotypic
- 16S rRNA sequence
- Other marker genes
- % DNA-DNA similarity
- % G+C composition
- Whole genome

Morphology

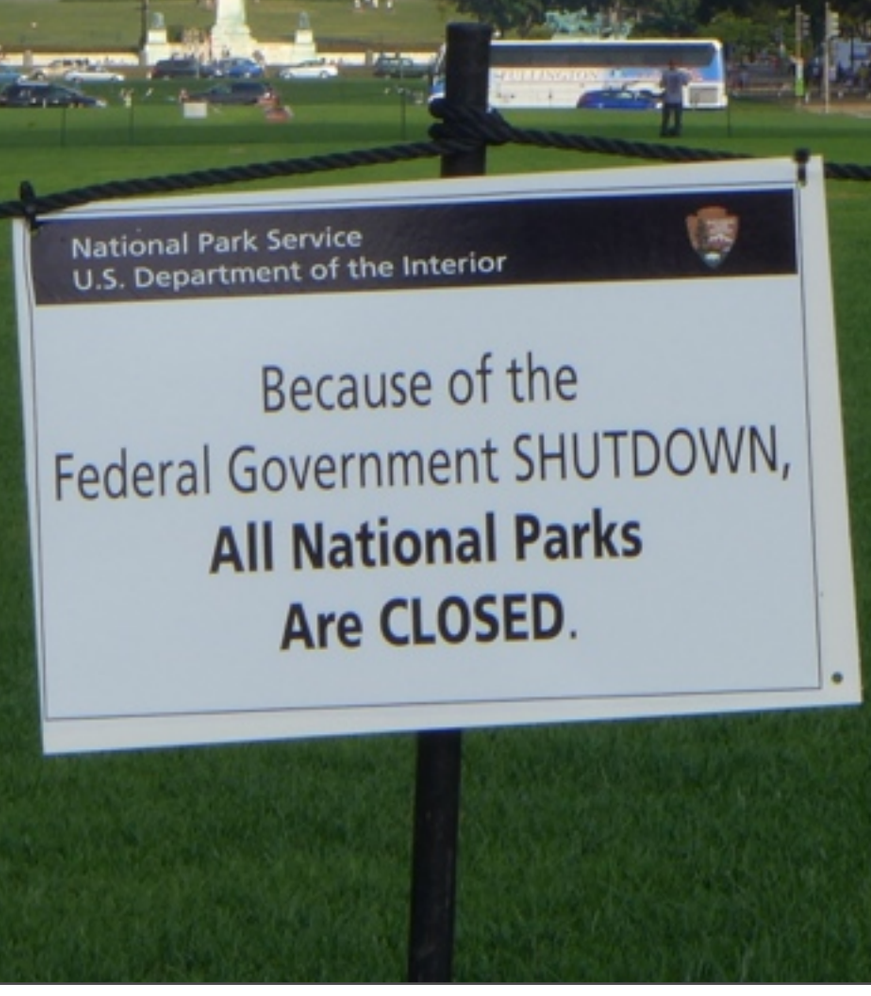
Micromorphology
- Cell size
- Cell shape
- Flagellation
- Sporulation
- Staining characteristics
- Other characteristics
- Intracellular inclusions
- Extracellular features
- Life cycle

Macromorphology
- Growth on solid surfaces
  - Colony morphology
- Growth in liquid
- Pigment production
- Other features

Chemotaxonomy
- Fatty acids
- Polar Lipids
- Mycolic Acids
- Respiratory quinones
- Peptidoglycan composition
- Polyamines

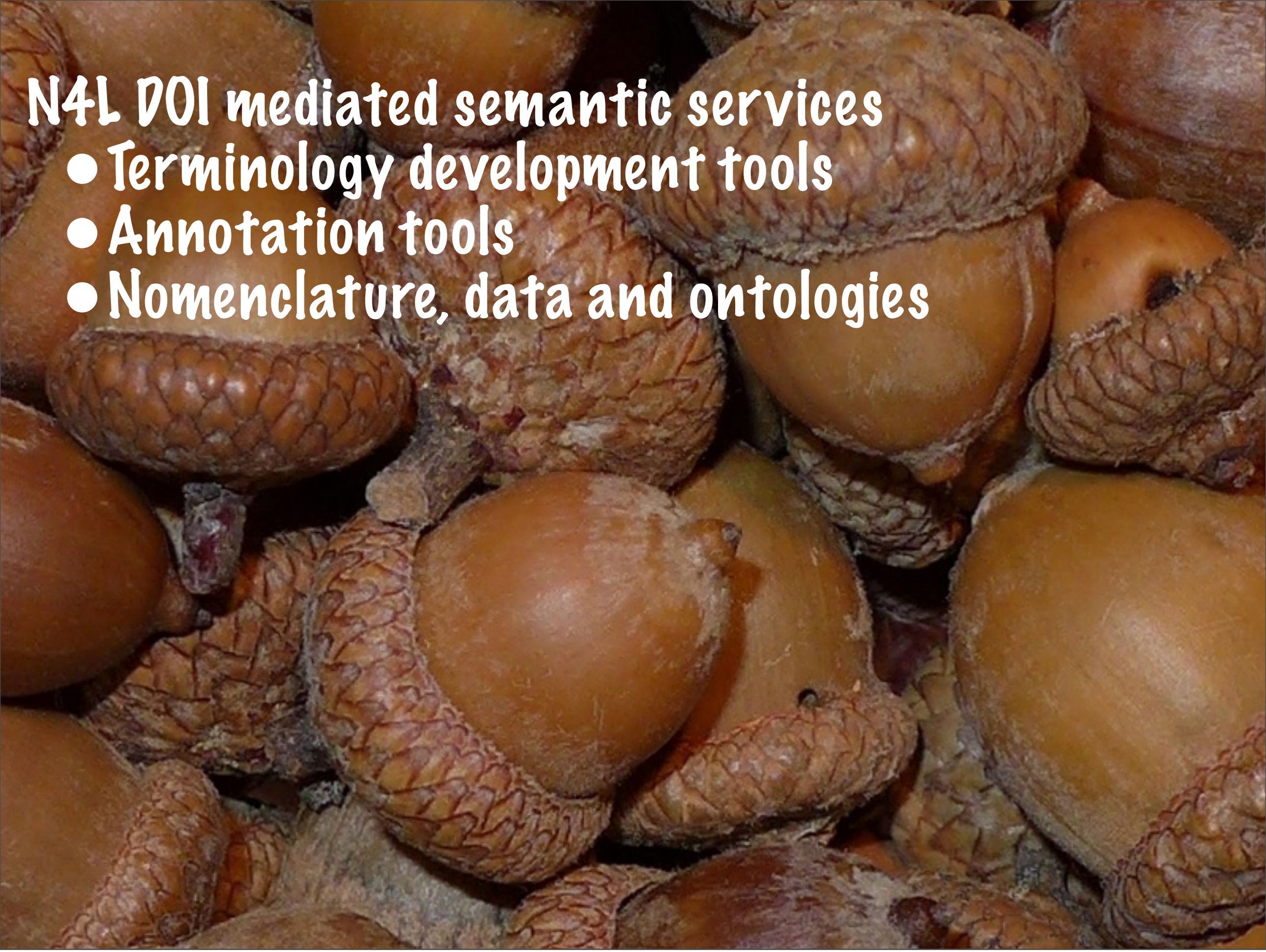
Physiological
- terminal e- acceptor
- substrate utilization
- metabolic end-products
- sensitivity/tolerance to
  - chemical and physical agents

Technical and non-technical barriers



N4L DOI mediated semantic services

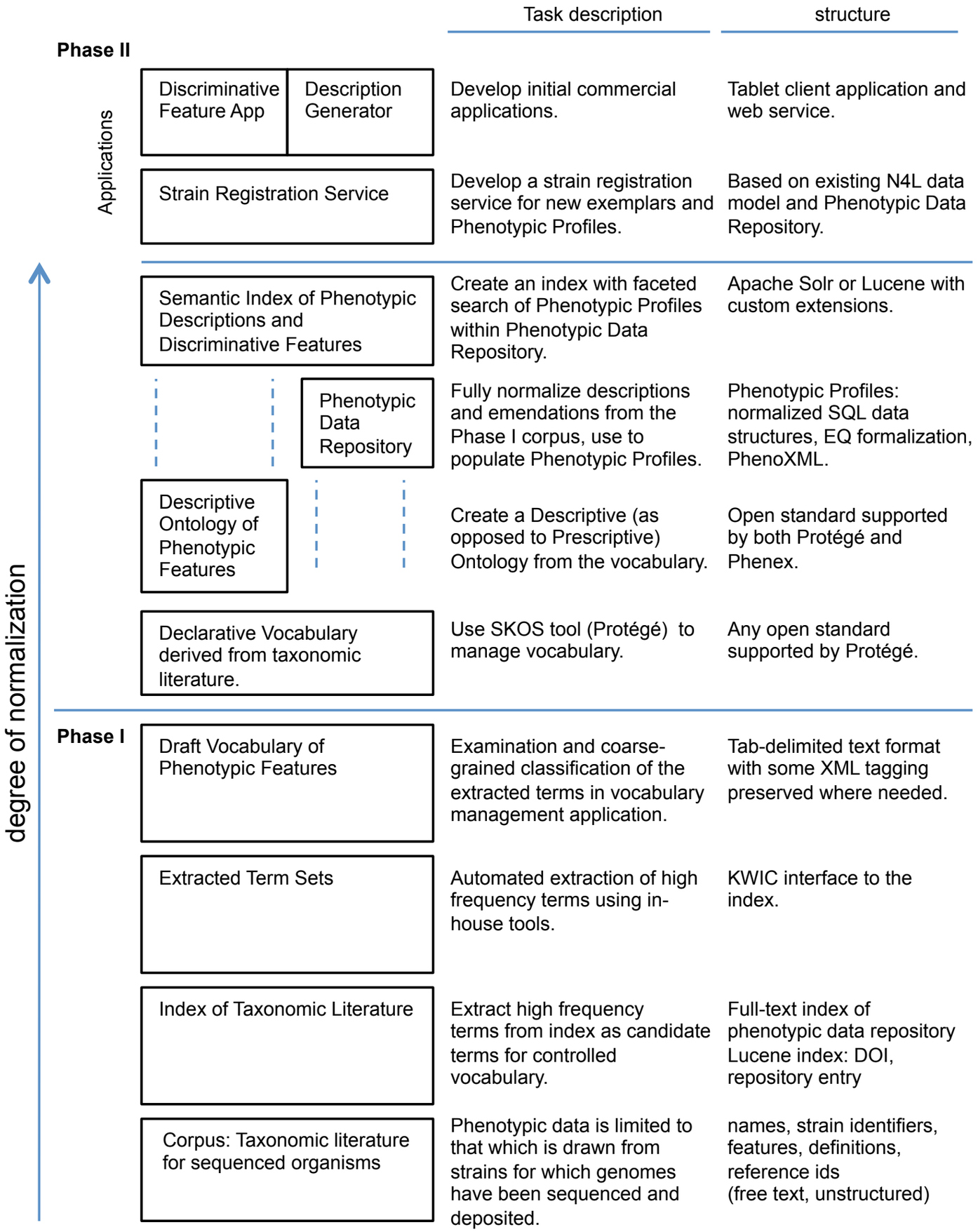
- Terminology development tools
- Annotation tools
- Nomenclature, data and ontologies



Project personnel

Chuck Parker
Nenad Krzavac
Kevin Petersen
Grace Rodrequez
Amber Roberts

Visit us at http://services.namesforlife.com



Phase I	Draft Vocabulary of Phenotypic Features	Examination and coarse-grained classification of the extracted terms in vocabulary management application.	Tab-delimited text format with some XML tagging preserved where needed.
	Extracted Term Sets	Automated extraction of high frequency terms using in-house tools.	KWIC interface to the index.
	Index of Taxonomic Literature	Extract high frequency terms from index as candidate terms for controlled vocabulary.	Full-text index of phenotypic data repository Lucene index: DOI, repository entry
	Corpus: Taxonomic literature for sequenced organisms	Phenotypic data is limited to that which is drawn from strains for which genomes have been sequenced and deposited.	names, strain identifiers, features, definitions, reference ids (free text, unstructured)

NamesforLife Semantic Index

Name Information

phylum class subclass order suborder family genus species subspecies

Collimonas

10.1601/nm.8548

Article Information

Int J Syst Evol Microbiol 2004; **54**(3):857.

10.1099/ijs.0.02920-0

Description of *Collimonas* gen. nov.

Cells are strictly aerobic, straight or slightly curved, Gram-negative rods, 0·3–0·5×1·0–2·0 μm. They occur singly and possess flagella (mostly one to three polar, but in some cases, several lateral) and pili when cultured in liquid media. Oxidase activity is positive; catalase activity is negative or weakly positive. Maximal growth is observed between 20 and 30 °C, without a sharp optimum. Maximum temperature that supports growth is approximately 35 °C. Optimal growth occurs at pH 6·5. Cells are able to hydrolyse colloidal chitin and milk proteins, but not lichenan or cellulose. Isolates vary in their ability to hydrolyse colloidal chitosan (deacetylated chitin). On minimal colloidal chitin/yeast agar, circular cleared haloes (final diameter, 4–10 mm) that support little translucent biomass are produced. Halo formation is completely repressed in chitin agar that contains TSB or glucose. On 0·1× TSB agar, colony morphology is variable. A wide range of sugars, alcohols, organic acids and amino acids can be metabolized. Several di- and trisaccharides cannot be used as substrates. In purified sand, cells proliferate upon introduction of intact, living fungal hyphae of various species of soil fungi. Major cellular fatty acids are $C_{16:0}$ and $C_{16:1\omega7c}$. DNA G+C content is 57–62 mol%. A member of the *β-Proteobacteria*, related most closely to the genera *Herbaspirillum* and *Janthinobacterium* in the family '*Oxalobacteraceae*', order '*Burkholderiales*'. Characteristics useful to differentiate the genus *Collimonas* from these related genera are given in Tables 1 and 2, as well as in the Results and Discussion. So far, the genus *Collimonas* is only known to occur in slightly acidic sandy dune soils. The type species is *Collimonas fungivorans*.

Etymology

Collimonas (Col.li.mo'nas. L. masc. n. *collis* hill; Gr. n. *monas* a unit, monad; N.L. fem. n. *Collimonas* cell from the hill).

KWIC Index : 714 occurrences of '"c16:1ω7c"'

haemolytic on sheep blood. The predominant fatty acid is **C16 : 1ω7c**. The G+C content is 64.5 mol% (sd 0.1) (det
talase is not produced. The major cellular fatty acid is **C16 : 1ω7c**. Cell-wall murein is type A4α containing l-
fungi. Major cellular fatty acids are C16 : 0 and **C16 : 1ω7c**. DNA G+C content is 57–62 mol%. A member of the β-
llow, flexirubin-like pigment and contain C16 : 1ω5c and **C16 : 1ω7c** as the major fatty acids. They do not hydro
acturonate or l-arabitol. Major cellular fatty acids are **C16 : 1ω7c** and C16 : 0. Q-8 is the predominant respira
acturonate or l-arabitol. Major cellular fatty acids are **C16 : 1ω7c** and C16 : 0. Q-8 is the predominant respira
acturonate or l-arabitol. Major cellular fatty acids are **C16 : 1ω7c** and C16 : 0. Q-8 is the predominant respira

Filters (inclusive) invert

- *Extracted Terms* — dictionary (4,763 terms)
- *Bacterial Names* — dictionary (15,346 terms)
- *Authorities* — dictionary (3,906 terms)
- *PubChem CID-MeSH* — dictionary (63,702 terms)
- *KEGG Lipids* — ontology (2,739 terms)
- *PubChem MeSH-Pharm* — dictionary (6,133 terms)
- *American National Corpus* — dictionary (237,267 terms)
- *Webster's 1934 Dictionary* — dictionary (310,536 terms)
- *Geographic Regions* — dictionary (24,734 terms)
- *Culture Collections* — dictionary (349 terms)
- *Nucleotides* — RegEx
- *Numbers* — RegEx
- *Term Length* — 3 ≤ length ≤ any
- *Term Frequency* — 5 ≤ frequency ≤ 1000

Unfiltered Terms (749)

term	frequency
c16 : 0	1375
iso-c15 : 0	979
c8	976
c4	965
c14	816
anteiso-c15 : 0	770
c16 : 1ω7c	714
iso-c16 : 0	663
c18 : 1ω7c	602
anteiso-c17 : 0	447
summed feature 3	419
iso-c15 : 0 2-oh	384
c18 : 0	375
c14 : 0	348
c15 : 0	271
iso-c17 : 0	263
iso-c17 : 0 3-oh	253
c18 : 1ω9c	239
iso-c14 : 0	236
c17 : 0	211
c17 : 1ω8c	153
iso-c17 : 1ω9c	140
c18 : 1	138
c12 : 0	123
c16 : 1	117
c12 : 0 3-oh	109
iso-c15 : 0 3-oh	99
c19 : 0 cyclo ω8c	97
c10 : 0 3-oh	91
summed feature 4	88
c16 : 1ω5c	86
c17 : 0 cyclo	80
c17 : 1ω6c	78
iso-c15 : 1	74
c16 : 1ω9c	69
c16 : 0 3-oh	64
i-c15 : 0	64
iso-c15 : 1 g	63
c20	62
c14 : 0 3-oh	61
10-methyl c18 : 0	59
11-methyl c18 : 1ω7c	58

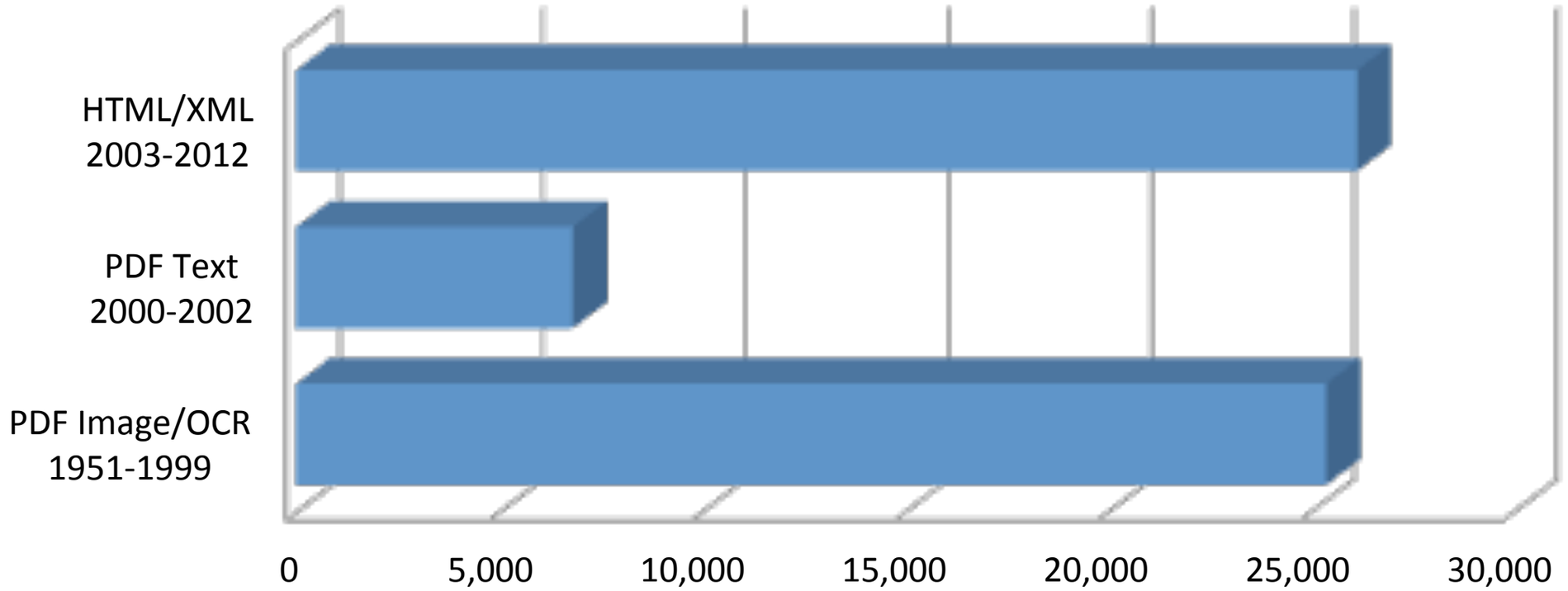
"c16 : 1ω7c"

Phase II

Applications

	Task description	structure
Discriminative Feature App / Description Generator	Develop initial commercial applications.	Tablet client application and web service.
Strain Registration Service	Develop a strain registration service for new exemplars and Phenotypic Profiles.	Based on existing N4L data model and Phenotypic Data Repository.
Semantic Index of Phenotypic Descriptions and Discriminative Features	Create an index with faceted search of Phenotypic Profiles within Phenotypic Data Repository.	Apache Solr or Lucene with custom extensions.
Phenotypic Data Repository	Fully normalize descriptions and emendations from the Phase I corpus, use to populate Phenotypic Profiles.	Phenotypic Profiles: normalized SQL data structures, EQ formalization, PhenoXML.
Descriptive Ontology of Phenotypic Features	Create a Descriptive (as opposed to Prescriptive) Ontology from the vocabulary.	Open standard supported by both Protégé and Phenex.
Declarative Vocabulary derived from taxonomic literature.	Use SKOS tool (Protégé) to manage vocabulary.	Any open standard supported by Protégé.

Scoring functions



The corpus

pre-2000 descriptions [corrected OCR, from 2,964 descriptions
2003-2009 freely available on web (3,338 articles)
2010-2012 provided via partnership with SGM (1,080 articles)

Yielded 5,750 descriptions and emendations

41.full.pdf - Adobe Reader

File Edit View Window Help

published

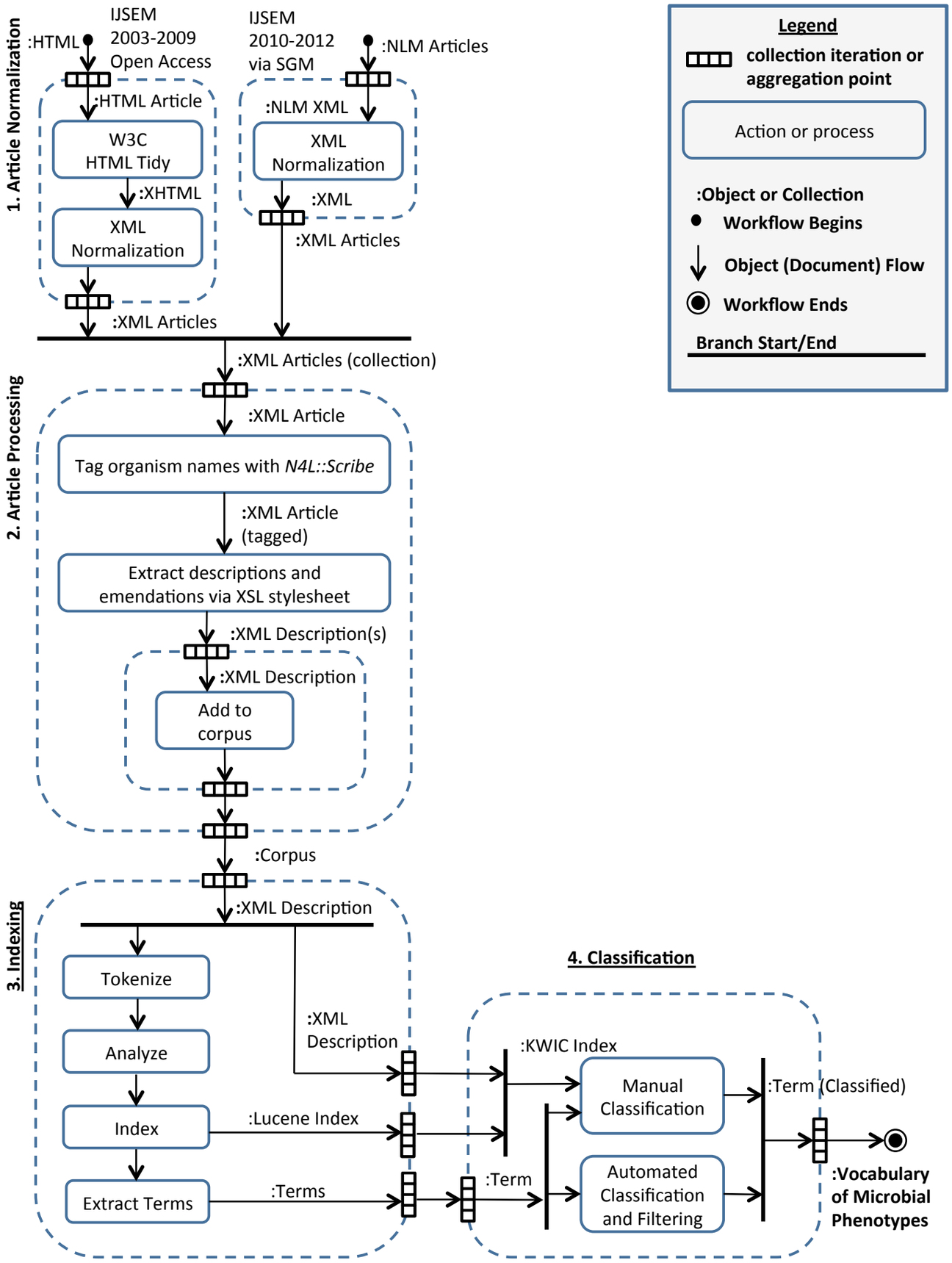
Obviously if *Pseudomonas*
Migula 1894 is to be recognized
as a genus it must be associated
with the names of the species
designated and described by
Mugula in 1895 Alternatively
the generic name *Pseudomonas*
might be recognized as having
been validly published in 1895
Migula (1900 p 884) (2) re-
cognized that the name *Bacterium*
aeruginosum Schroter 1872 (3)
antedated *Bacillus pyocyaneus*
Gessard 1882 (4) and published
the name as *Pseudomonas aeruginosa*
(Schröter) Migula 1900.
Buchanan (1918 p 18) desig-

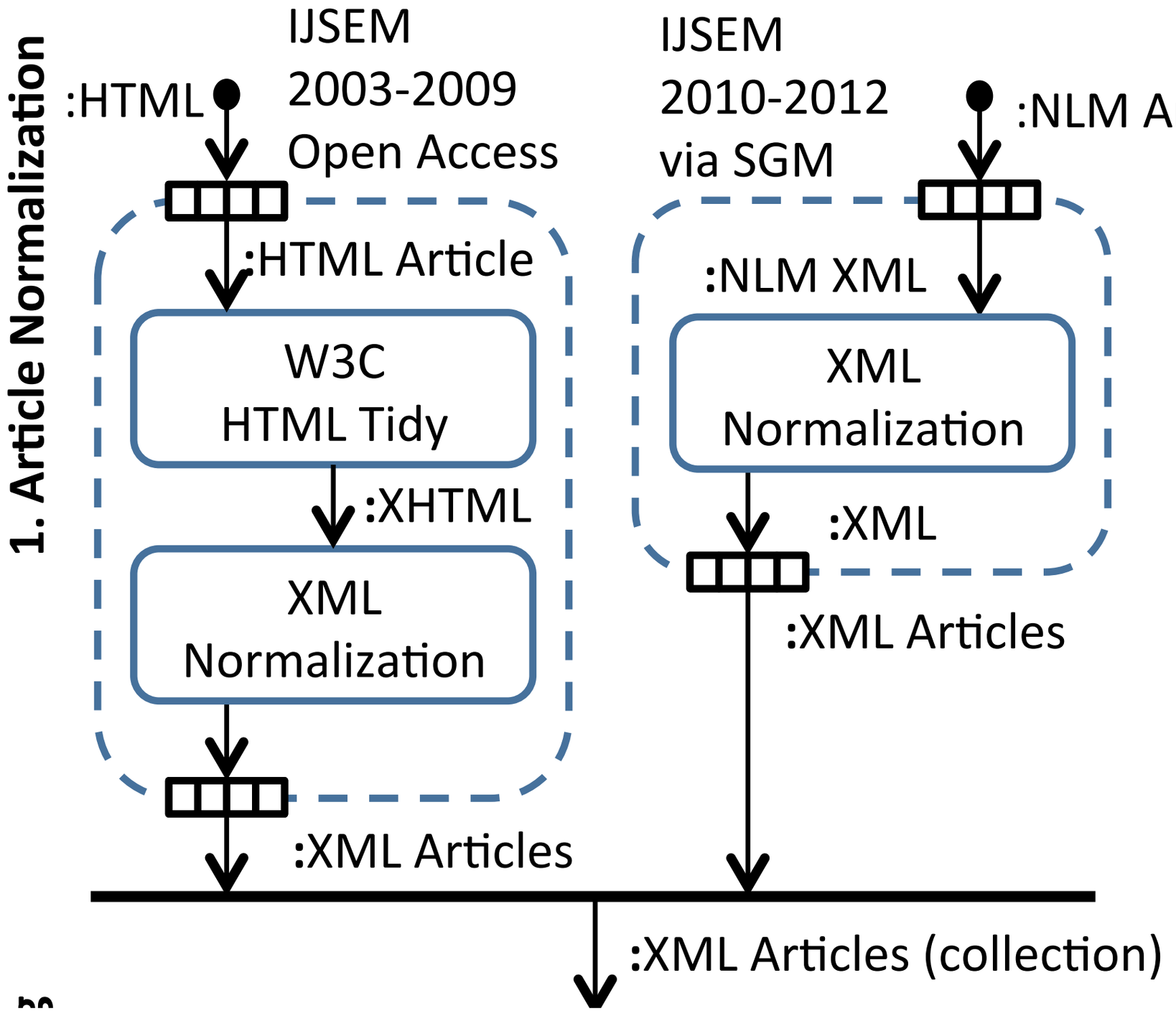
6.50 x 8.50 in

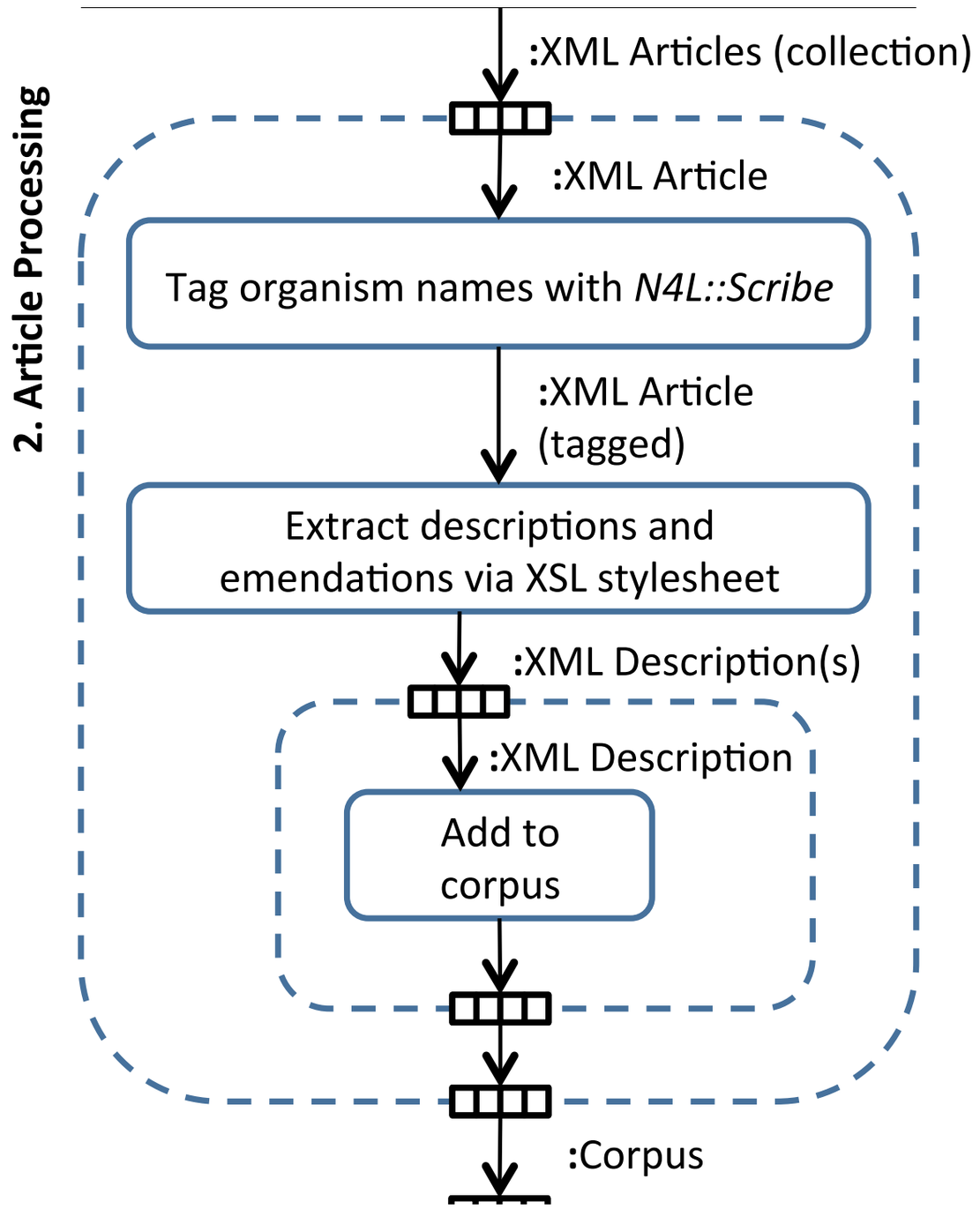
IJSB_1_41.txt - Notepad

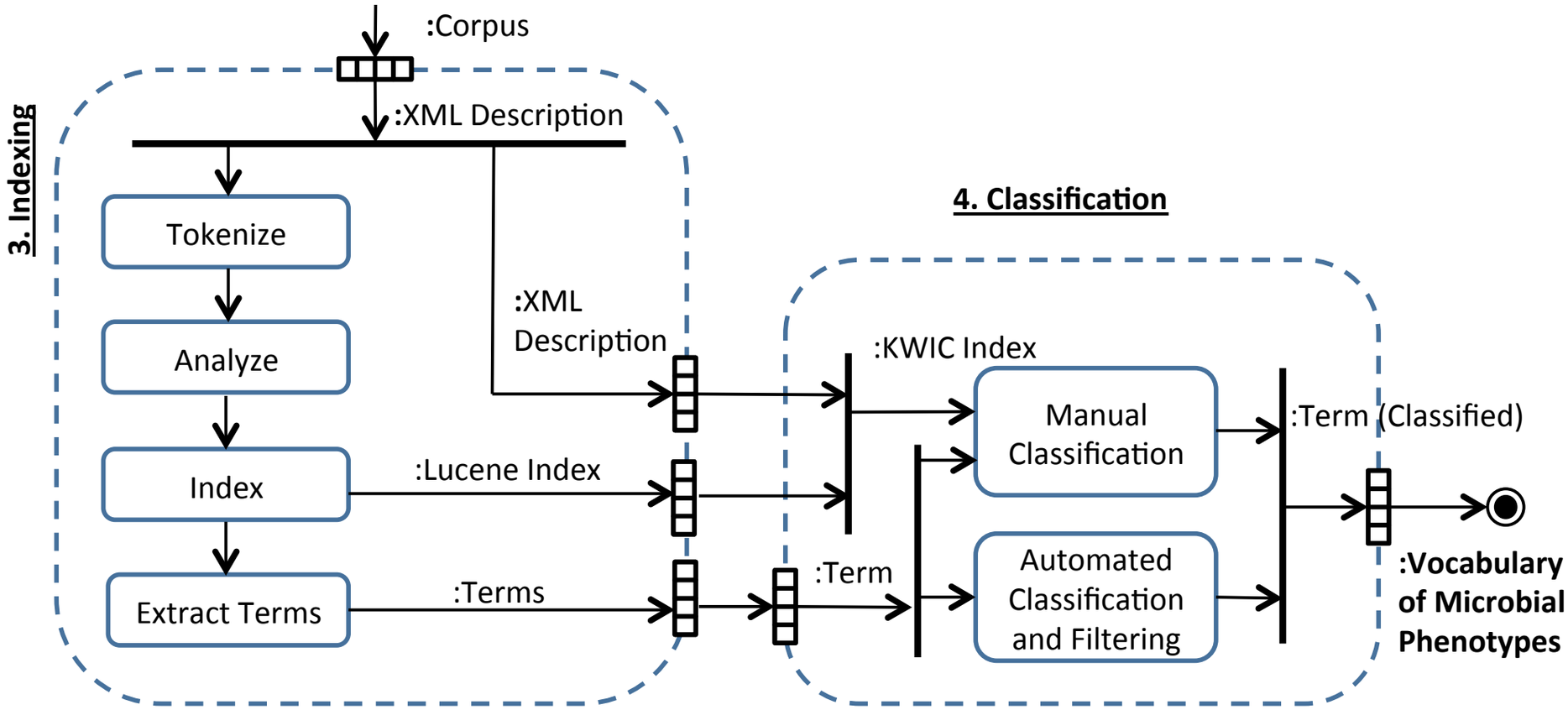
File Edit Format View Help

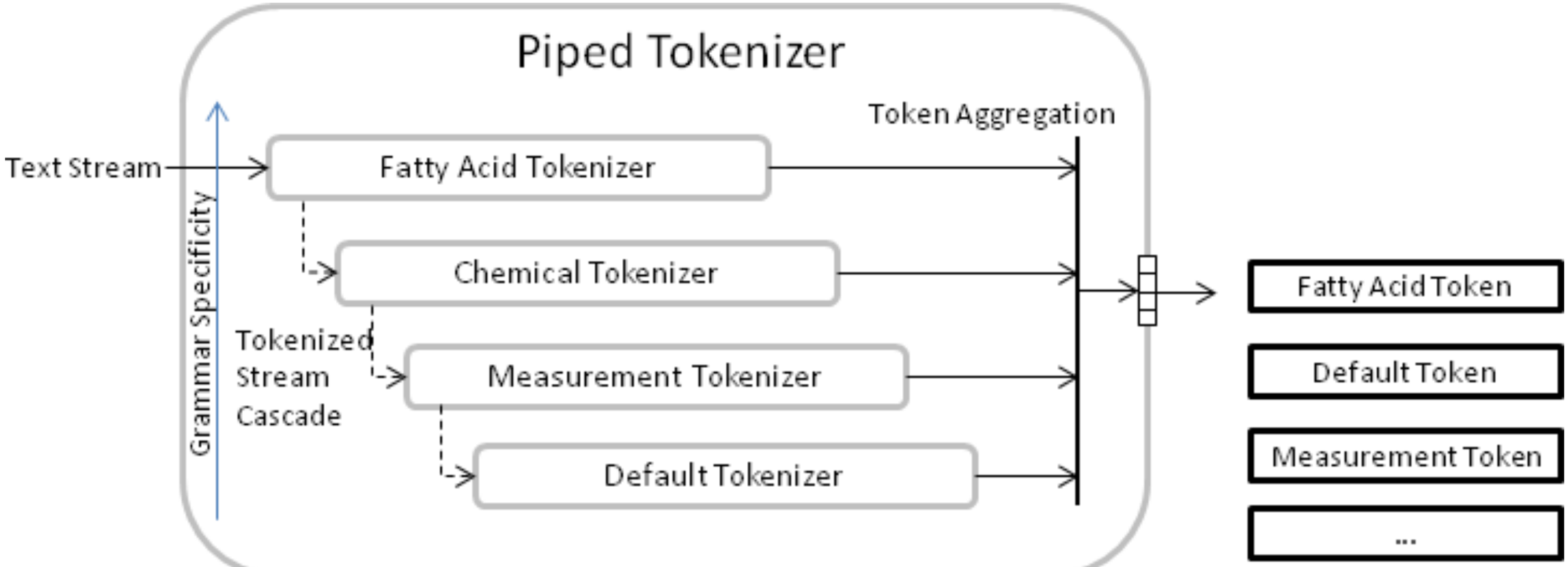
Obviously i f
Migula 1894 IS t o be recognized
as a genus i t must be associated
with t h e names of the species
designated and described by
Mugula i n 1895 Alternatively
might be recognized as havin
k e n valid!y published i n 188s
the genesic name f t!~aC/7il~"%(l-
Gessard 1882 (4) and published
(1) schrEter J Schizomycetes
i n C o b . F Kryptogamcn FloTa
von Sc73lesien 3 136 174, Qe6
(2) Migula. W Systein der
Baktesien 2 884 -3uc
6 urch Bacterien gebildete P i g









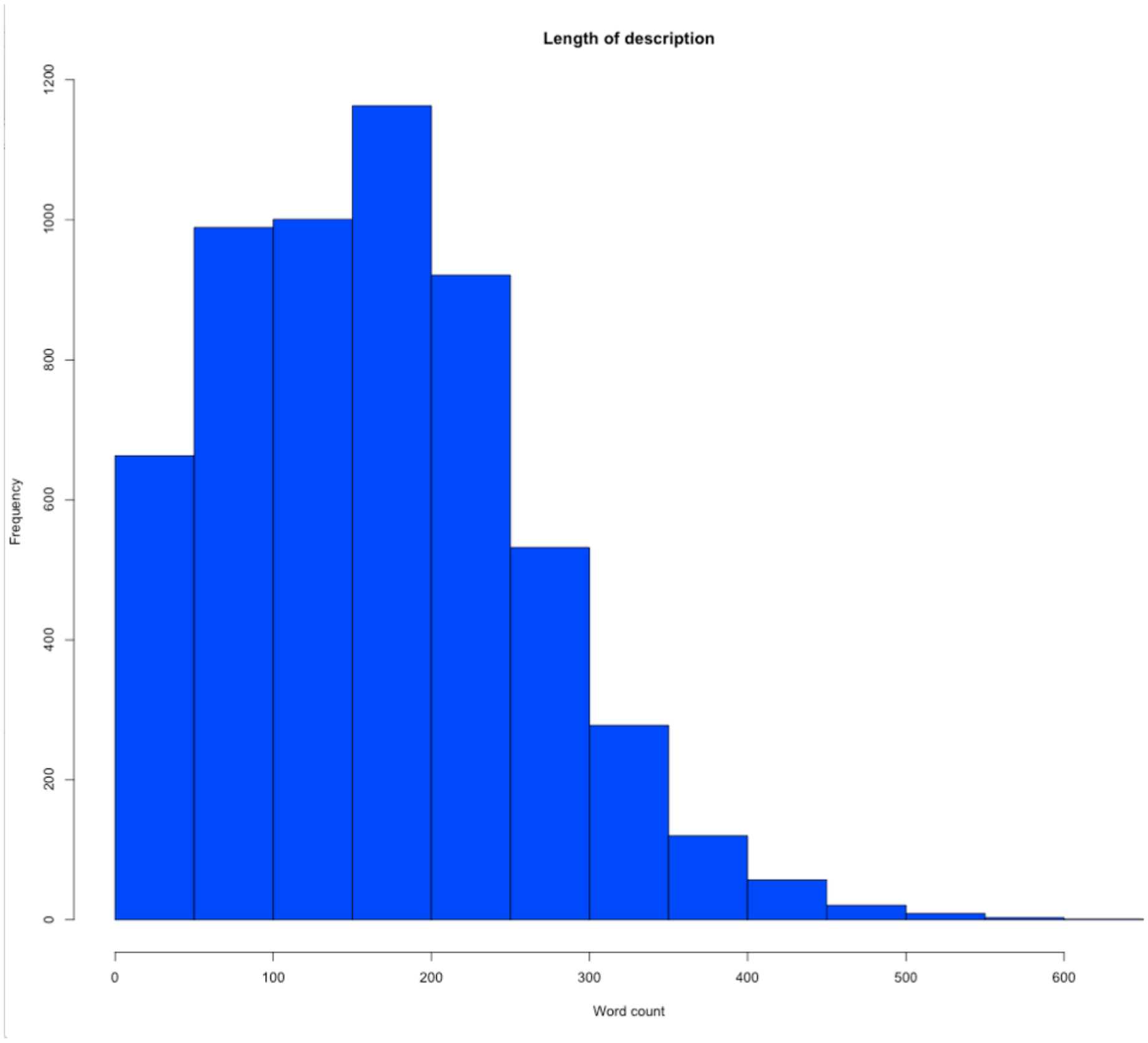


Taxonomic composition of the corpus

Rank	Frequency
phylum	4
class	17
subclass	4
order	31
suborder	19
family	131
genus	1095
species	4382
subspecies	75

Corpus metrics

Phylum	n	total	mean	Phylum	n	total	mean
Proteobacteria	2220	272688	123	Chlorobi	15	1691	113
Actinobacteria	1255	152022	121	Synergistetes	15	951	63
Firmicutes	994	120469	121	Planctomycetes	13	1604	123
Bacteroidetes	765	105384	138	Deferribacteres	12	803	67
Euryarchaeota	157	16990	108	Armatimonadetes	11	459	42
Verrucomicrobia	48	5499	115	Caldiserica	6	219	37
Deinococcus-Thermus	38	4594	121	Fusobacteria	6	488	81
Chloroflexi	34	1845	54	Gemmatimonadetes	6	208	35
Aquificae	28	1953	70	Thermodesulfobacteria	5	423	85
Tenericutes	27	2118	78	Nitrospirae	4	323	81
Spirochaetes	24	2309	96	Chlamydiae	3	277	92
Crenarchaeota	18	1416	79	Cyanobacteria	2	92	46
Acidobacteria	16	1305	82	Lentisphaerae	2	64	32
Thermotogae	16	1626	102	Chrysiogenetes	1	75	75



Not all descriptions are equally informative

Description of *Kribbella koreensis* (Lee *et al*. 2000) comb. nov.

Kribbella koreensis (ko.re.en′sis. N.L. fem. adj. *koreensis* pertaining to Korea, the location of the soil sample from which the type strain was isolated).

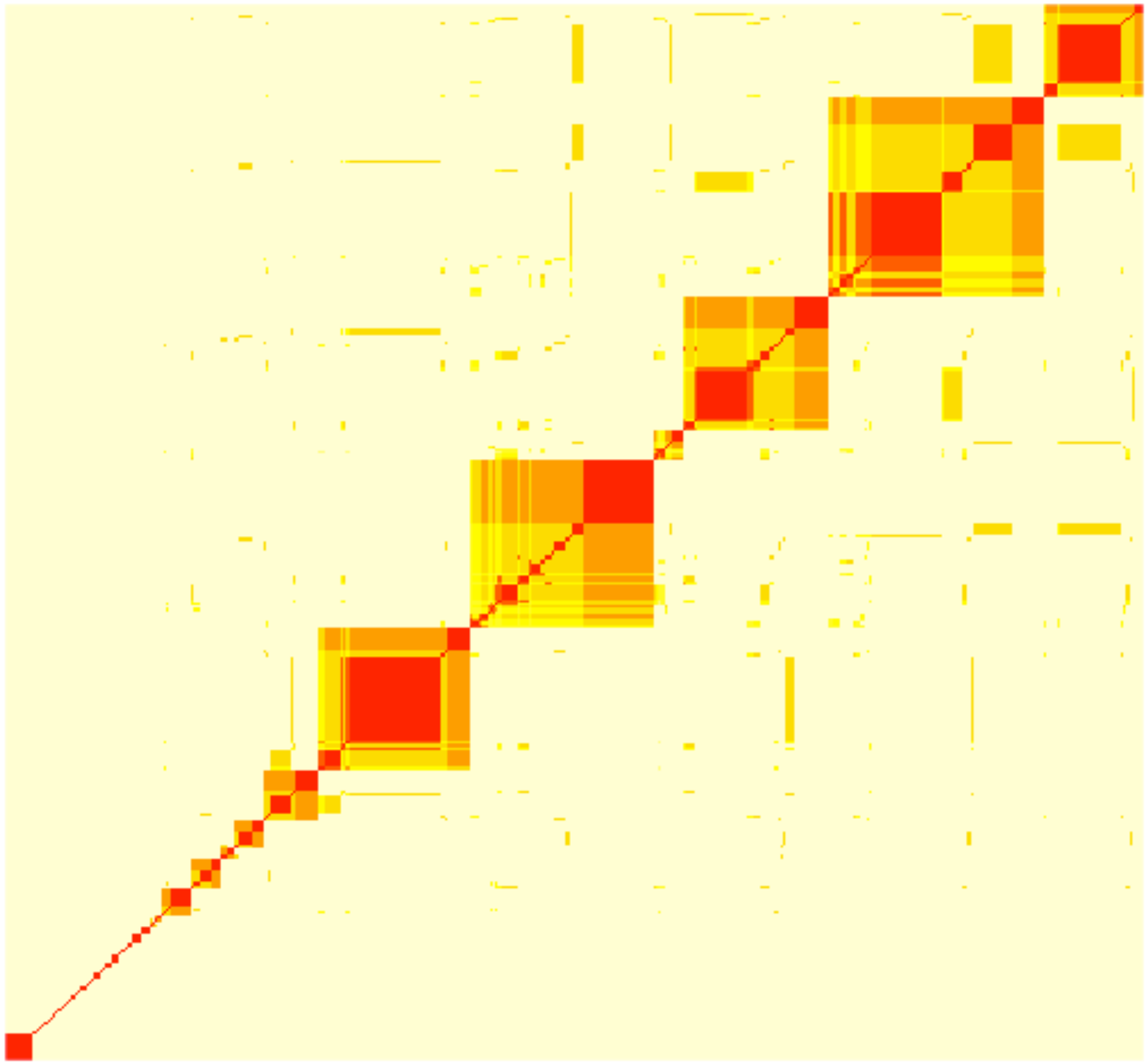
The description is identical to that of Lee *et al*. (2000).

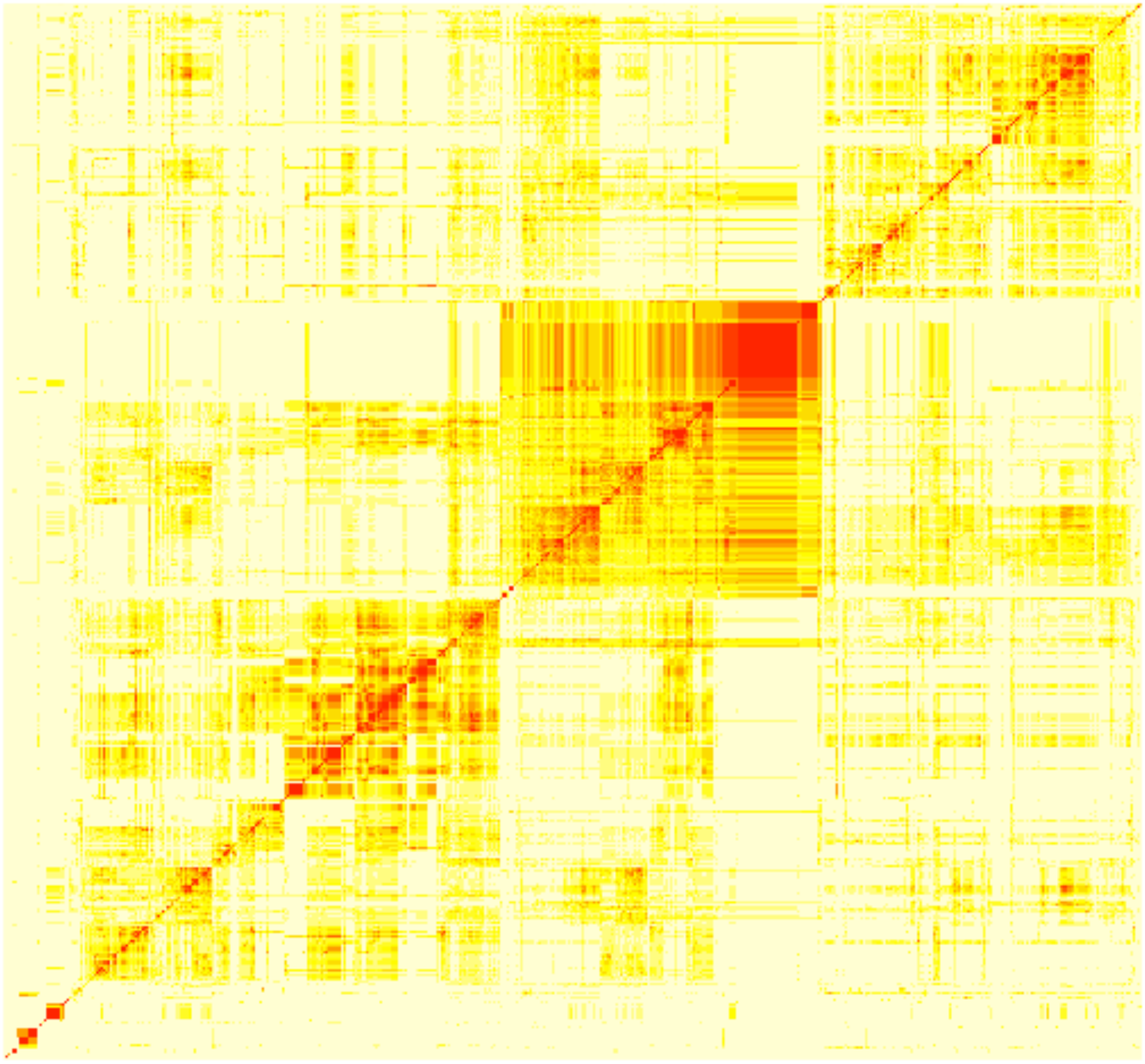
Emended description of *Anoxybacillus pushchinoensis* corrig. Pikuta *et al*. 2000

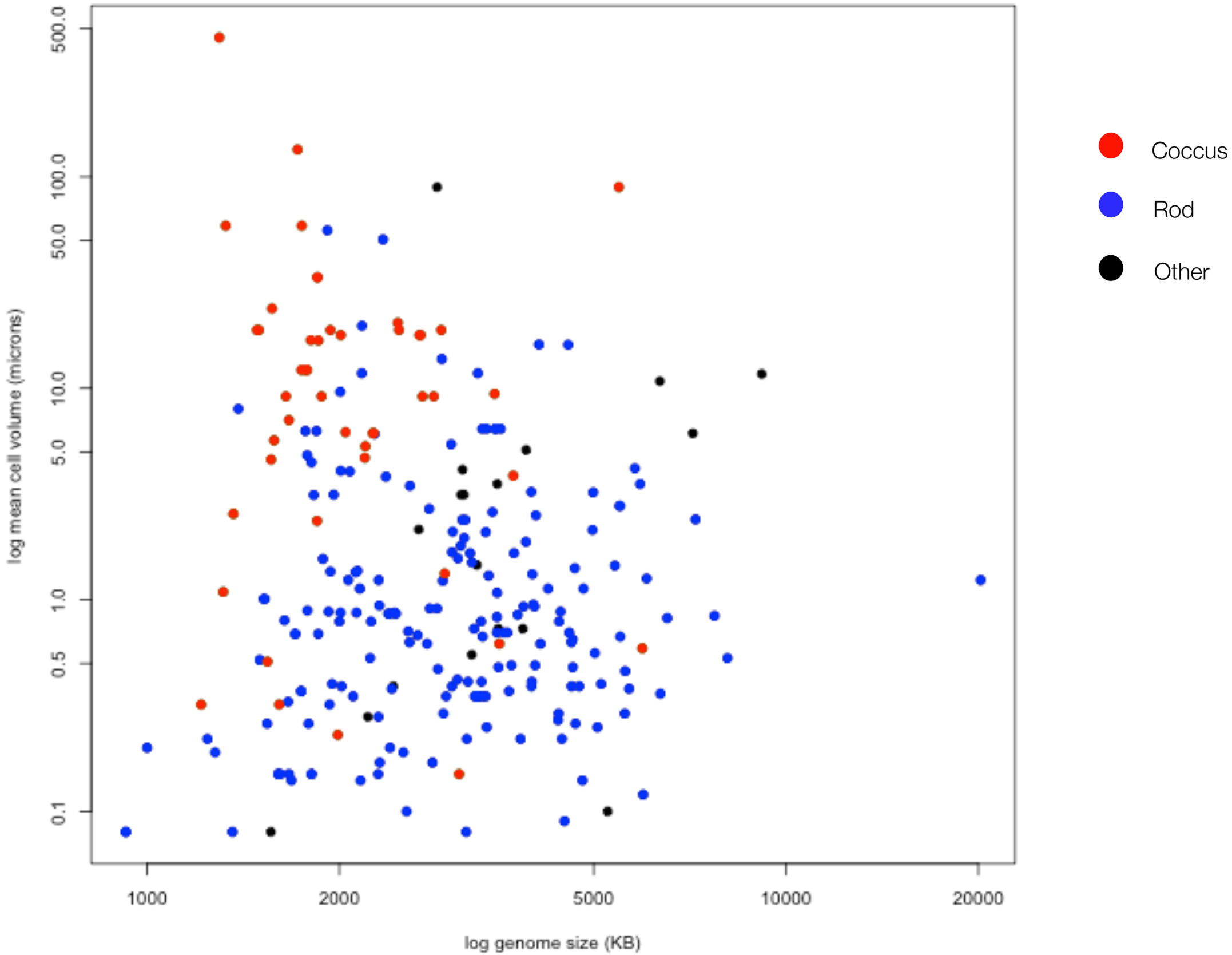
Aerotolerant anaerobe.

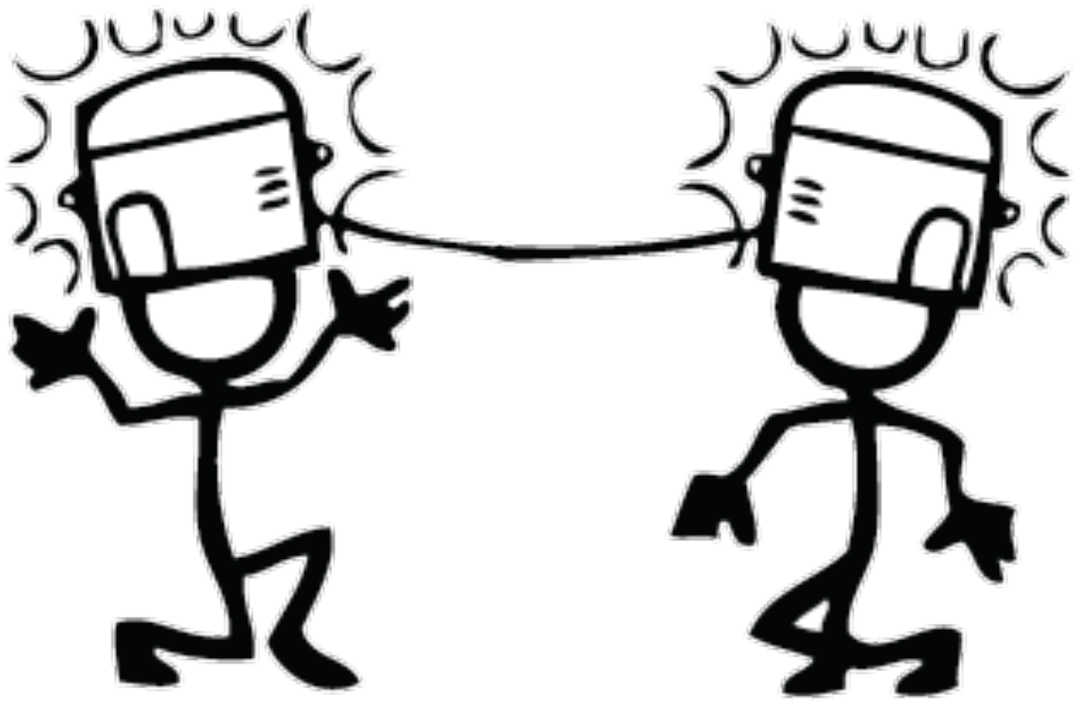
Fatty Acid	Extracted
Tokenized	733
Reduced	675
Synonyms	412
Straight chain	399
Unsaturated	
mono	255
di	12
tri	4
tetra-hexa	6
branched	315
hydroxy	141
DMA	13
cyclo	4

Matched in LipidMaps	205
Unidentified	529
Reduced set	675
Normalized	412
synonyms (11-21)	6
synonyms (2-9)	76
unique	330







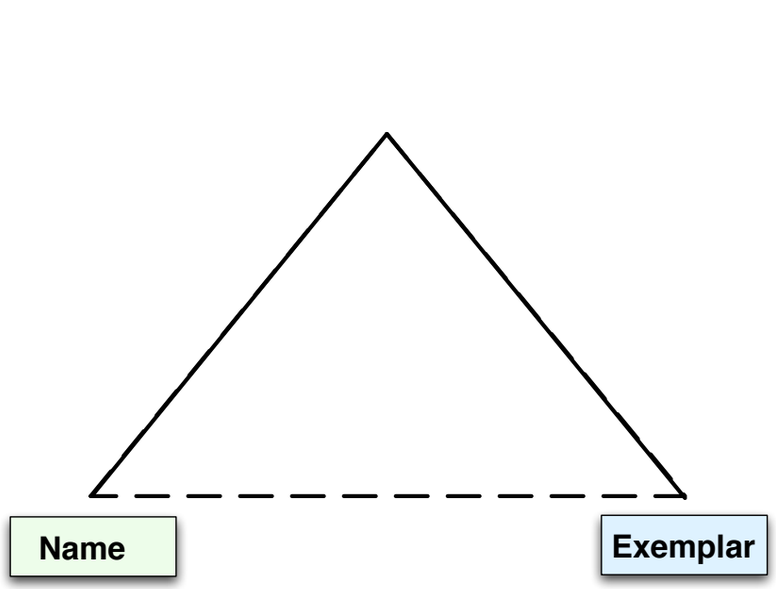


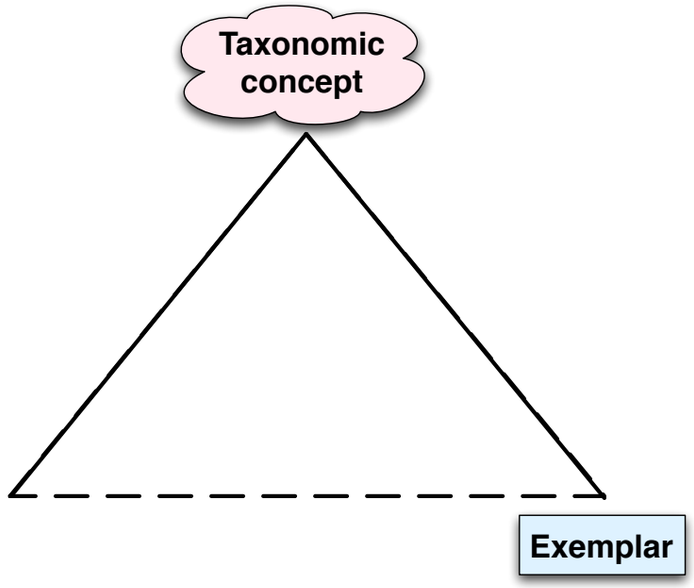
Our goal

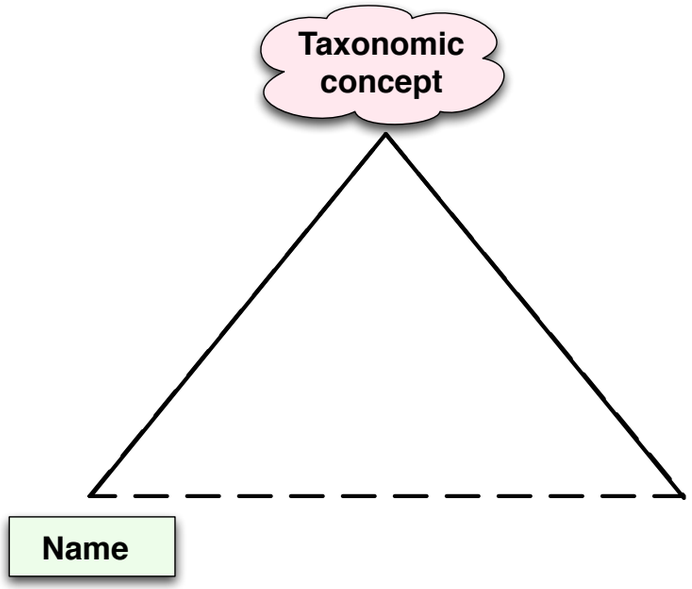
To provide on-demand access to the correct information so that end-users can interpret like a subject matter expert.

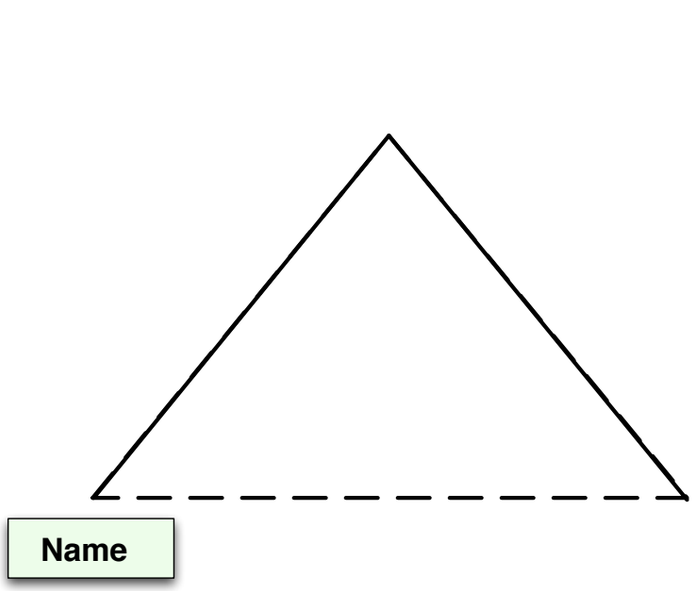
Following up on Folker's bright idea

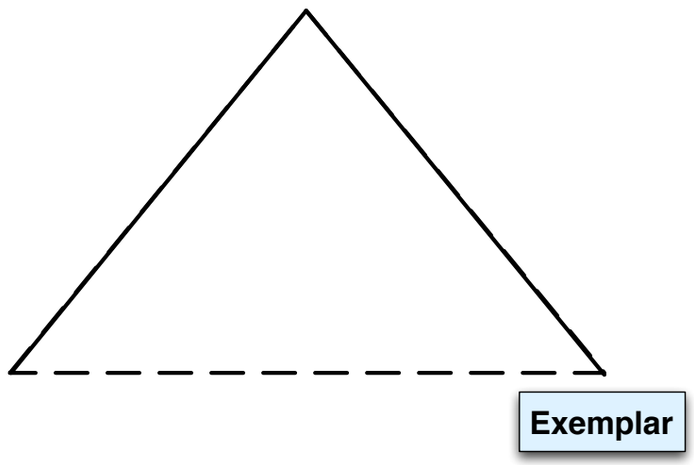
The reality of modern microbiology

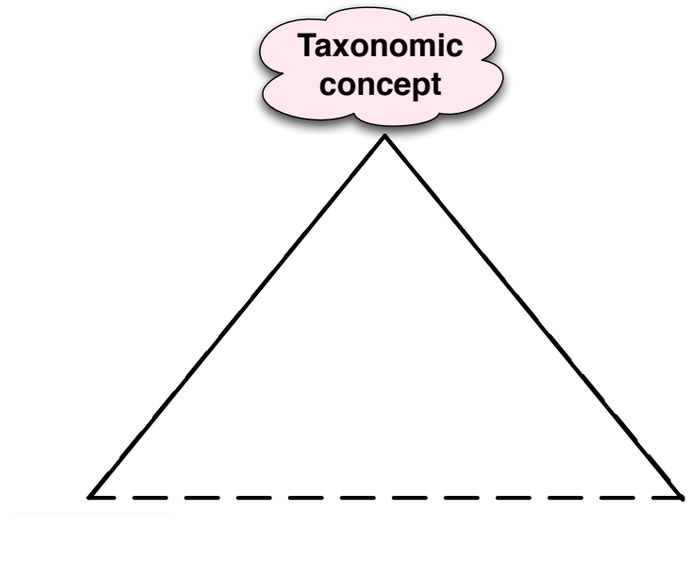












Phase II

			Task description	structure
Applications	Discriminative Feature App	Description Generator	Develop initial commercial applications.	Tablet client application and web service.
	Strain Registration Service		Develop a strain registration service for new exemplars and Phenotypic Profiles.	Based on existing N4L data model and Phenotypic Data Repository.

The parts are falling into place, so what would it take?

NamesforLife™
Bringing meaning to life ...

Catherine Lyons
Charles T. Parker
Dorothea Taylor
Sarah Wigley
Nicole Osier
Kara Mannor
Grace Rodreguez

SGM

Ron Fraser
Robin Dunford
Karen Rowlett

MICHIGAN STATE UNIVERSITY

James R. Cole
Jordan Fish
Xiong Wang
Donna MacGarell

INTERNATIONAL COMMITTEE ON SYSTEMATICS OF PROKARYOTES

Brian Tindall
Aharon Oren

ATCC™

Timothy G. Lilburn

University of Nebraska Lincoln

Khalid Sayood
Ufuk Nalbantoglu
Sam Way

The Microbial Earth Project

Nikos Kyrpides
Hans-Peter Klenk

SIGS

Oranmiyan Nelson

DTU

Dave Ussery

Beta-testers

Thanks to our partners, collaborators and sponsors

eXtyles®

Fairview Research
Harmonizing Business and Technology

TREPAREL
INFORMATION SOLUTIONS

doi>
TM

SGM

INTERNATIONAL COMMITTEE ON SYSTEMATICS OF PROKARYOTES

The Microbial Earth Project

muci
MICHIGAN UNIVERSITIES COMMERCIALIZATION INITIATIVE

U.S. DEPARTMENT OF ENERGY | Office of Science

MICHIGAN
ECONOMIC DEVELOPMENT CORPORATION

Your logo could go here

3rd-party apps

Microbial Earth | *SciVerse* App | *eXtyles* | *Treparel* Extension

Application Layer

N4L Guide | *N4L Scribe* | NamesforLife Abstracts | Alert Services | Search | Report

Service Layer

The service layer can be exposed to partners for developing custom applications or services.

N4L Guide Toolkit (client API)

N4L Scribe SOAP API

Curation, Production and Publishing Workflow

AJAX API

SOAP API

Business Logic Layer

N4L Guide Server

Semiotic Analysis / Clustering

Document Indexer (Lucene)

Data Access Layer

Retrieval API | N4L Semantic Tagger | NamesforLife Contextual Index | Local Index (Private)

nomenclature snapshot | N4L API | Alexandria API | Custom API

Database Layer

At the database layer, we have a federated database model, where each database is kept up to date by one of our partners. NamesforLife is responsible for the nomenclature.

Genomes OnLine Database

N4L Nomenclature Database

N4L Corpus IJSEM, ASM, SAM, PMC, SIGS, HighWire Press

Alexandria Patent Database

Local Document Corpus and library content.