Who? What? Where?
Public databases for nucleotide sequence data (NSD)





Dr. Amber Hartman Scholz, Deputy to the Director

Acknowledgments

Fabian Rohden
Germany

Sixing Huang
Germany / China

Gabriele Dröge
Germany

Contributing authors (alphabetical):

Katharine Barker, USA
Walter G. Berendsohn, Germany
Jonathan A. Coddington, USA
Manuela da Silva, Brazil
Jörg Overmann, Germany
Ole Seberg, Denmark
Michelle van der Bank, South Africa
Xun Xu, China

KEG6
AHS3

Why do scientists put NSD in a public database for everyone to see and use?

1. <u>Publishing</u>: Journals *will not publish* your paper unless the data is in the INSDC. (Regardless of nationality.)

2. <u>Funding</u>: Grant agencies very often *require* it.

3. <u>Ethics</u>: Scientific reproducibility, integrity, data security

Slide 3

KEG6 I think this is important but would see this more as part of the discussion. Can we park this slide at the end and use it in the discussion if needed?
Karger, Elizabeth GIZ; 25/10/2019

AHS3 Please trust my judgment that this is essential for understanding how and why the database system is the way it is.
I have discussed with Marcel and he does not cover this aspect at all and is much more focused on use and definition.
This slide explains why the architecture has been established.
Amber H. Scholz; 31/10/2019

Where do scientists put their NGS?
How d

The Complete Ge Haloferax volcanii DS2 pla IV3, complete sequence i DS2, a
Model Archa

Amber L. Hartman, Cédri
Hoda K
[view a
Publish
Artic
Abstr
Introd
Resu
Mate
Supp
Ackn
Authc
Refer
Read
Media
Figur

Category
Display Name (*)
NCBI Taxon ID (*)
Domain (*)
Phylum (*)
Class (*)
Order
Family
Genus (*)
Species (*)
Strain (*)
Culture Collection ID
Biosafety Level
Project Information
Project description
Accession #s
Project Type ( )
Project Status (*)
Contact Name (*)
Contact Email (*)
GC Percent
Sequencing Center Nam
Sequencing Center url (
Funding Agency Name
Funding Agency url
Publication Journal
Publication Volume
Publication link (url)
Sequencing Informat
Sequencing Status (*)

KEG7
KEG10
KEG3
AHS1

D13378.1.1470, Haloferax gibbonsii, 1470 nuc
EU308225.1.1438, Haloferax sp. FB247_8, 1438 nuc
AB081732.1.1473, Haloferax lucentense, 1473 nuc
KY827082.1.1265, Haloferax sp., 1265 nuc
DQ458842.1.1429, Haloferax sp. YT226, 1429 nuc
CP011947.2862162.2863642, Haloferax gibbonsii, 1481 nuc
AB037474.1.1472, Haloferax alexandrinus, 14
AOLL01000014.72909.74350, Haloferax
AOLJ01000012.98386.99827, Haloferax
AB663378.1.1473, Haloferax lucentense, 1473
HQ438275.1.925, Haloferax sp. PR13, 925
GQ478070.1.1448, Haloferax sp. CT4-2, 1448
GQ478076.1.1448, Haloferax sp. CS1-9, 1448
FJ746722.1.1473, Haloferax sp. H4, 1473 nuc
AY425724.1.1404, Haloferax volcanii, 1404 nu
CP001956.2770163.2771635, Haloferax volcan
CP001956.1598192.1599664, Haloferax volcan
AOHU01000104.40516.41975, Haloferax volca
AB074566.1.1472, Haloferax volcanii, 1472 nuc
KC354391.1.933, Haloferax
KF650666.1.925, Haloferax v
KC354398.1.90
KC354382.1.927, Haloferax l
KC354384.1.931, Haloferax vol
KC354396.1.934, Haloferax
KC354394.1.946, Haloferax sp. B
DQ915828.1.901, Haloferax vo
KC354392.1.928, Haloferax vo
KX197500.1.1415, Haloferax sp., 141
KF321798.1.1379, Haloferax sp. A2FP3cs, 1379 nuc
JX067386.1.1435, Haloferax sp. C27, 1435 nuc
JF781313.1.903, Haloferax sp. KPS1, 903 nuc
D14128.1.1469, Haloferax denitrificans, 1469 nuc
(99 of 127 seqs polyphyletic)

silva high quality ribosomal RNA databases
Home SILVAngs Browser Search ACT Download Do

SILVA

Welcome to the SILVA rRNA database project
A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package ARB.

For more background information → Click here

SILVAngs

silvangs
Check out our service for Next Generation Amplicon data

SILVA Alignment, Classification and Tree (ACT) Service
The SILVA ACT service combines alignment, search and classify as well as reconstruction of trees in a single web application.

SILVA ACT is available at: → www.arb-silva.de/act

SILVA Tree Viewer
The SILVA Tree Viewer is a web application to browse and query the SILVA guide trees.

A technical preview is available at www.arb-silva.de/treeviewer

https://today.uconn.edu/school-stories/papke-archaea/#

Slide 4

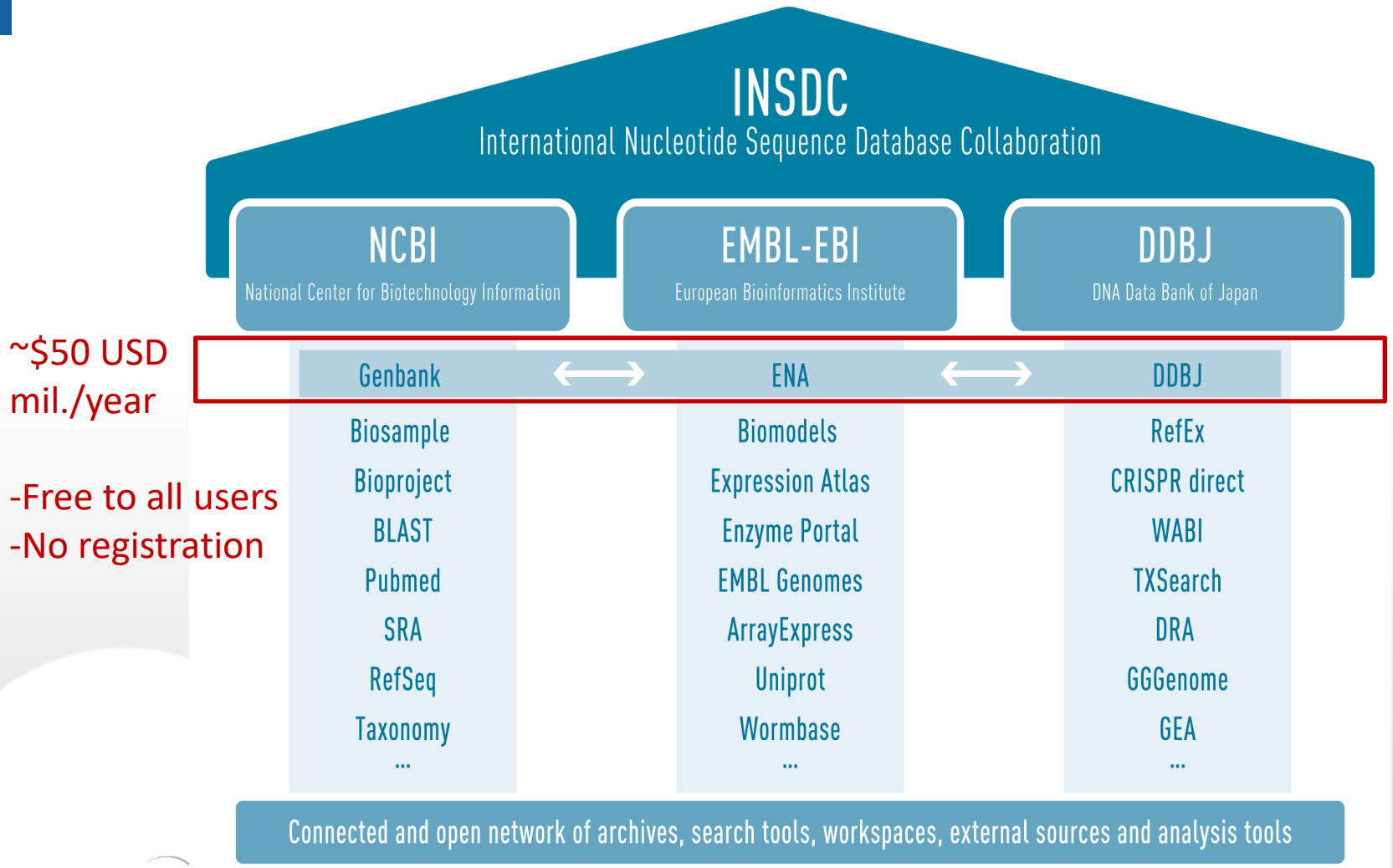
KEG3 NSD needs a brief explanation - you need to point out that this may not only be nucleotide sequences, e.g.proteins
Karger, Elizabeth GIZ; 25/10/2019

KEG7 Karger, Elizabeth GIZ; 25/10/2019

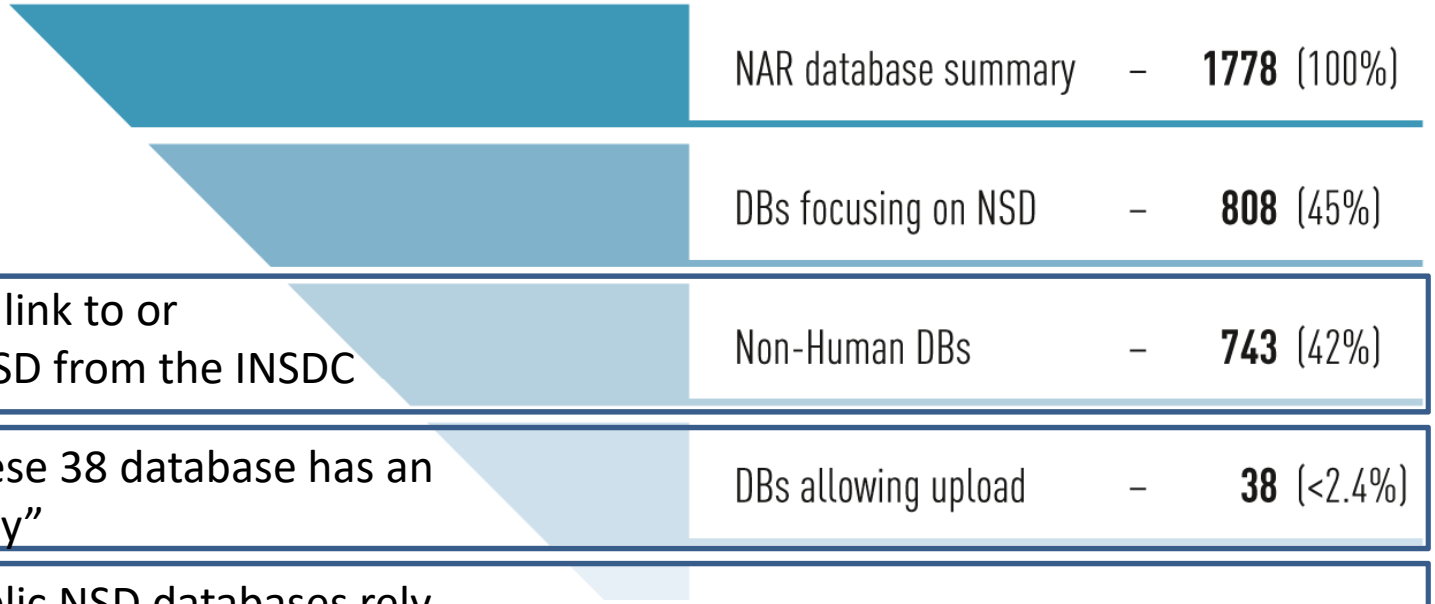
KEG10 can we just use data or sequences in the title?
Karger, Elizabeth GIZ; 25/10/2019

AHS1 This is ONLY NSD. The research presented here is only on NSD. Many lessons learned here can be extrapolated to protein sequence data but they are rarely or only partially relevant to other forms of DSI
Amber H. Scholz; 27/10/2019

The INSDC is the core infrastructure but there are dozens of databases & tools in these organizations



How many public NSD databases are there? How many rely on the INSDC?



95% directly link to or download NSD from the INSDC

Only 1 of these 38 database has an "access policy"

100% of public NSD databases rely on the INSDC

"The S/MAR transaction database is free for users from non-profit organizations only. Users from commercial enterprises have to license, please contact marketing@biobase.de for details."

→ Access to this database does not require registration.

olicy (published 2002)

Science's COMPASS

LETTERS SCIENCE & SOCIETY POLICY FORUM BOOKS *ET AL.* PERSPECTIVES

Nucleotide Sequence Database Policies

1 THE INTERNATIONAL NUCLEOTIDE SEQUENCE Databases (INSD) has been an international collaboration between DDBJ, EMBL, and GenBank for over 14 years. Its advisory board, the International Advisory Committee, is made up of members of each of

2 the databases' advisory bodies. At their last meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSD, which is stated below.

Image not available for online use.

3 Individuals submitting data to the international sequence databases managed collaboratively by DDBJ, EMBL, and Gen-

4 Bank should be aware of the following:

1) The INSD has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.

2) The INSD will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.

3) All database records submitted to the INSD will remain permanently accessible maintained by the INSD is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.

5) Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.

The INSD is an outstanding example of success in building an immensely valuable, widely used public resource through voluntary cooperation across the international scientific community. This success has been achieved by following the guidelines and principles outlined above.

SOREN BRUNAK,[1*] ANTOINE DANCHIN,[2*] MASAHIRA HATTORI,[3†] HARUKI NAKAMURA,[4†] KAZUO SHINOZAKI,[5†] TARA MATISE,[6‡] DAPHNE PREUSS[7‡]

[1]Center for Biological Sequence Analysis, Technical University of Denmark, DK-2800 Lyngby, Denmark. [2]Genetics of Bacterial Genomes, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France. [3]Laboratory of Genome Information, Kitasato Institute for Life Science, Kitasato University, 1-15-1, Kitasato, Sagamihara, Kanagawa, 228-8555 Japan. [4]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita 565-0871, Osaka, Japan. [5]Laboratory of Plant Molecular Biology, RIKEN, 3-1-1 Koyadai, Tsukuba, Ibaraki, 305-0074 Japan. [6]Department of Genetics, Rutgers University, 604 Allison Road, Piscataway, NJ 08854–8082, USA. [7]Howard Hughes Medical Institute, University of Chicago, 1103 E. 57th Street, Chicago, IL 60637, USA.

*EMBL advisors.
†DDBJ advisors.
‡GenBank advisors.

ed access to all of the data records their databases

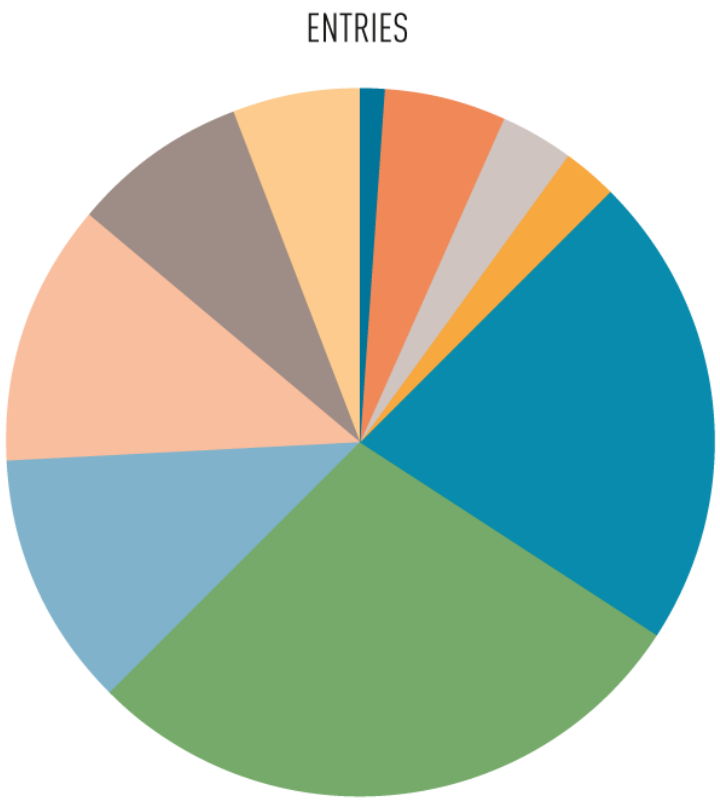
records that restrict access to the data, limit the use
prohibit certain types of publications based on these
s or licensing requirements will be included in any

INSD will remain permanently accessible as part of

es maintained by the INSD is **fully disclosed to the**

Science 298 (5597): 1333 15 Nov 2002

“ Who“ is in the public databases?



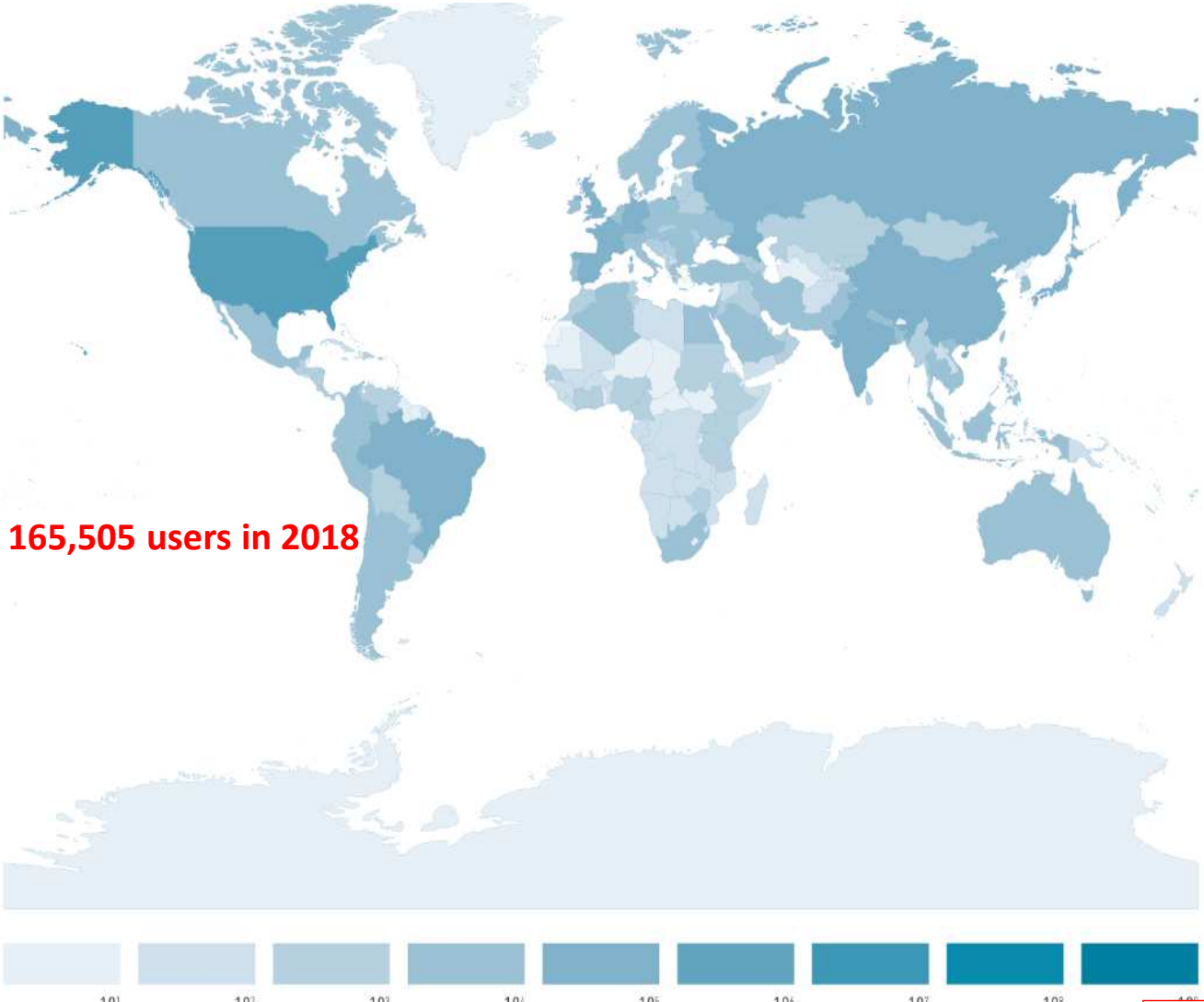
TAXONOMIC GROUP	ENTRIES	BASE PAIRS
Viruses	1,21%	0,32%
Bacteria/Archaea	5,60%	7,37%
Fungi	3,29%	1,95%
Single-celled Eukaryotes	2,56%	0,91&
Plants & Algae	21,61%	17,12%
Animals	28,45%	55,62%
Human	11,52%	6,25%
Model Organisms	11,93%	8,67%
Other/synthetic (microorganisms)	8,03%	1,16%
Unidentified (no tax. ID)	5,82%	0,61%

There are 10-15 million total users of INSDC. They live in every country in the world.

1.	United States	22.69 %
2.	China	15.42 %
3.	India	6.16 %
4.	Japan	3.97 %
5.	Germany	3.67 %
6.	United Kingdom	3.45 %
7.	France	2.84 %
8.	Brazil	2.83 %
9.	Spain	2.31 %
10.	Russian Federation	2.25 %

Costs: $3-5 per user

50% of users live in countries that do not contribute to NSD infrastructure costs



KEG9

Slide 9

KEG9 point out that this is a logarithmic scale - people might miss this

Karger, Elizabeth GIZ; 25/10/2019

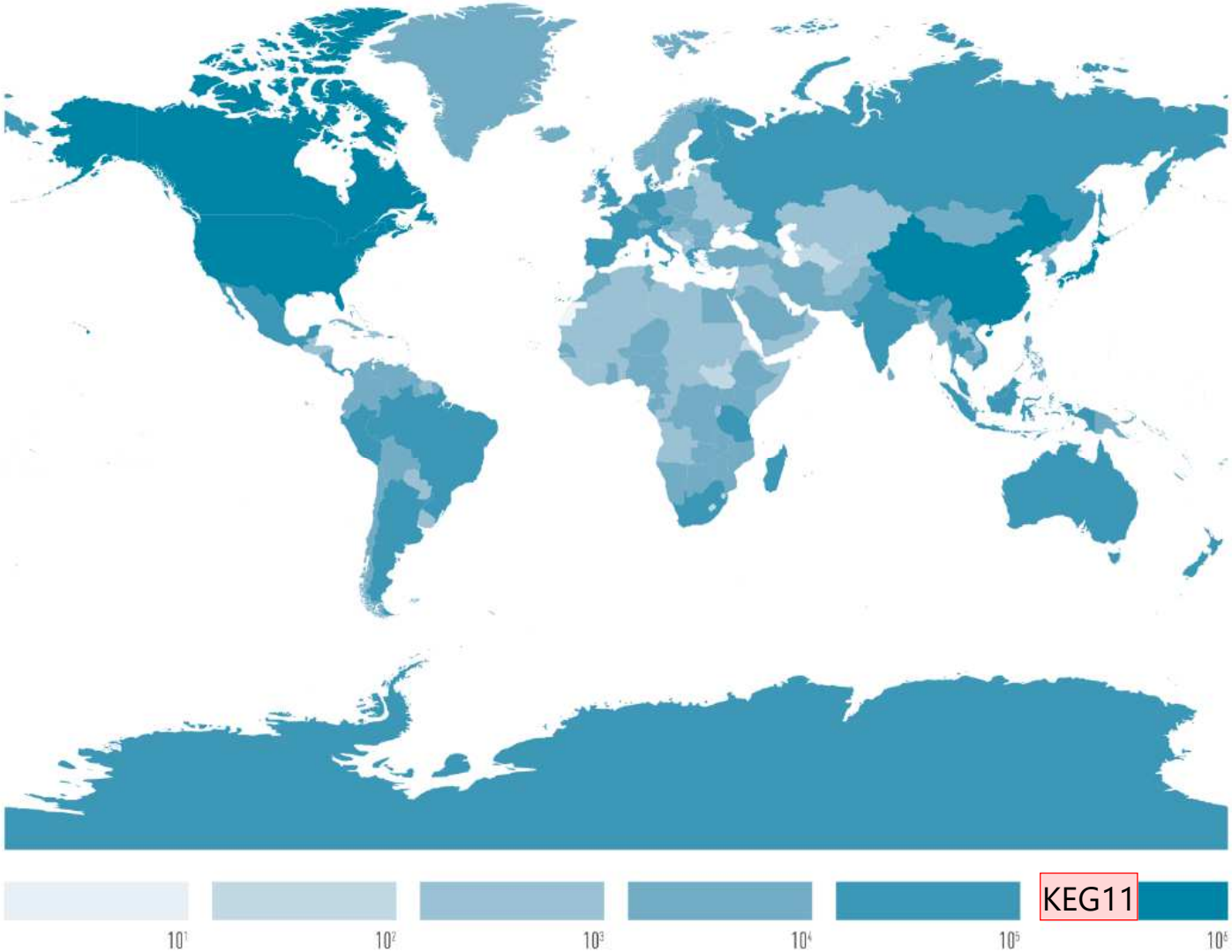
Where does the original GR f[illegible]hese NSD come from?

KEG5 KEG8 AHS5

What is the country of origin for non-human NSD?

1.	China	18.23 %
2.	United States	17.39 %
3.	Canada	9.10 %
4.	Japan	7.24 %
5.	India	3.46 %
6.	Australia	2.66 %
7.	Mexico	2.54 %
8.	Brazil	2.30 %
9.	Germany	1.83 %
10.	Spain	1.58 %

52% of NSD comes from 4 countries



Slide 10

KEG5 is this a little confusing? we need to clearly distinguish between where the sequences are are uploaded from vs where the organisms come from.
Karger, Elizabeth GIZ; 25/10/2019

KEG8 it would be interesting to have a slide about where the GR originally come from. Do you have this?
Karger, Elizabeth GIZ; 25/10/2019

KEG11 point out that this is a log scale
Karger, Elizabeth GIZ; 25/10/2019

AHS5 This map IS showing where the original GR was sourced from.

It is NOT showing where the sequences were produced/sequenced or where they were uploaded from. That data is unfortunately not machine-readable.
Amber H. Scholz; 31/10/2019

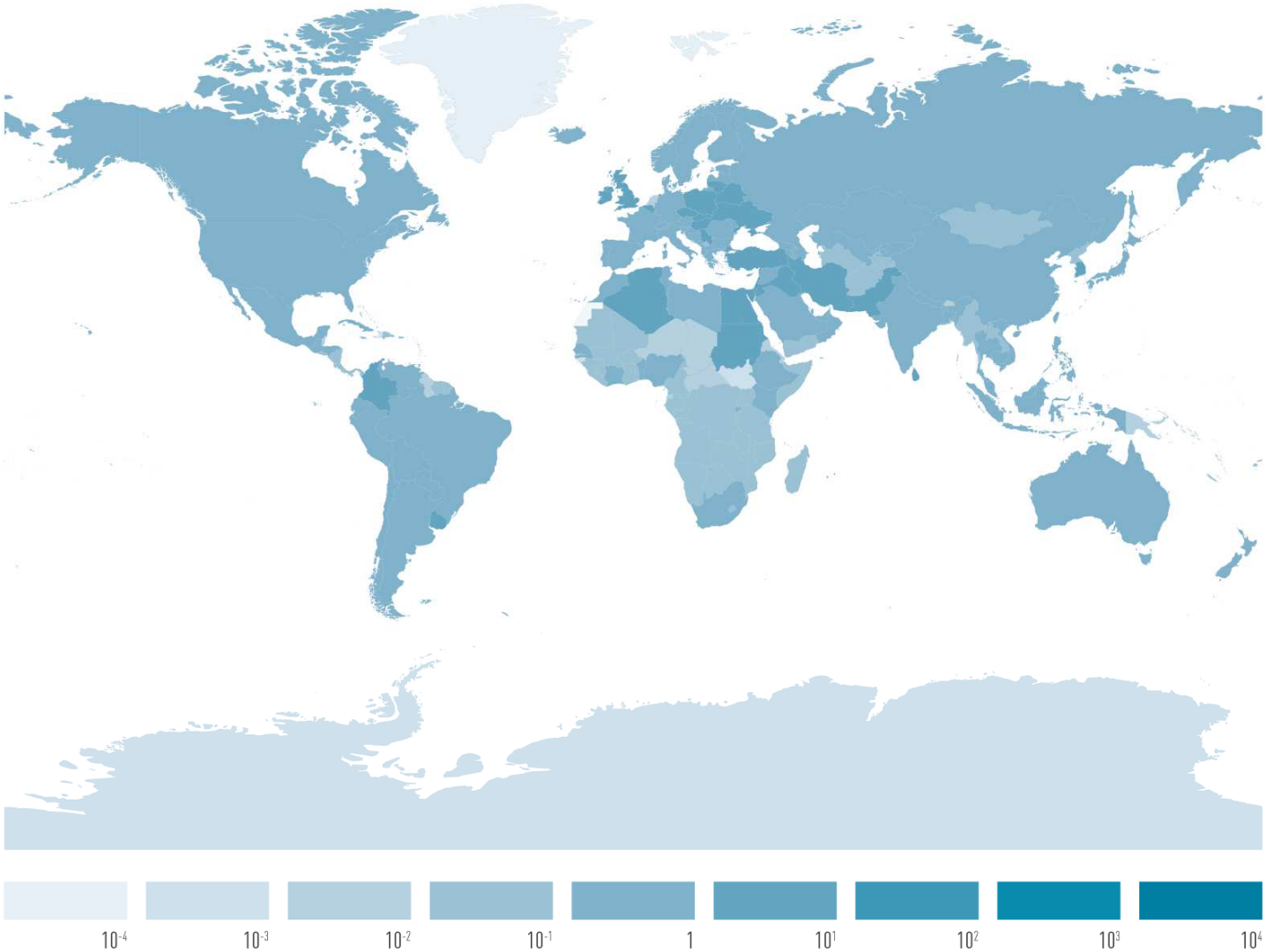
For each country: How many users are there per sequence entry in the INSDC?

How does database usage compare to provided sequences?

1.	Lebanon	27.68
2.	Ukraine	23.84
3.	Belarus	22.88
4.	Iraq	21.41
5.	Colombia	19.79
6.	Algeria	19.54
7.	Pakistan	19.37
8.	Republic of Korea	18.81
9.	Belgium	17.83
10.	Poland	17.47

Brazil's rank = 73 (0.32 users/seq)

USA rank = 71 (0.31 users/seq)



Take-home messages

1. Scientists cannot publish without submitting their NSD to the INSDC.
2. The INSDC will remain the core infrastructure for NSD (because of human/model organism NSD and geographical origin).
3. Access policies for public NSD databases are overwhelmingly open access.
4. NSD *comes* from the entire world, is *used* by the entire world, but is dominated by ~15 countries.

A Brazilian strain in our public catalog

Burkholderia sp.

DSM 103188

BACTERIA	ⓘ How to read the following data (Example)
Name:	***Burkholderia* sp.**
DSM No.:	**103188, Type strain**
Strain designation:	89
Isolated from:	native grassland soil
Country:	Brazil Southern Brazil, Municipality of São Joaquim, Santa Catarina
Date of sampling:	05.04.2011
Nagoya Protocol Restrictions:	Customers who purchase this strain from DSMZ should partner with a Brazilian institution and are free to conduct research and development activities, however before publishing results, applying for any intellectual property rights and / or marketing products, the Brazilian institution will have to register the R&D in the SisGen system. The user, in association with a Brazilian institution can conduct R&D with the material and the Brazilian partner must register the research in SisGen before the following cases: 1. Request of any intellectual property right 2. Commercialization of any intermediate product 3. Release of results, final or partial, in scientific or communication circles 4. Notification of finished product or reproductive material developed as a result of the access Note: This is in accordance with article 12 of Brazilian Law 13.123/2015.

Before user purchases a strain, she must commit to CBD/Nagoya restrictions

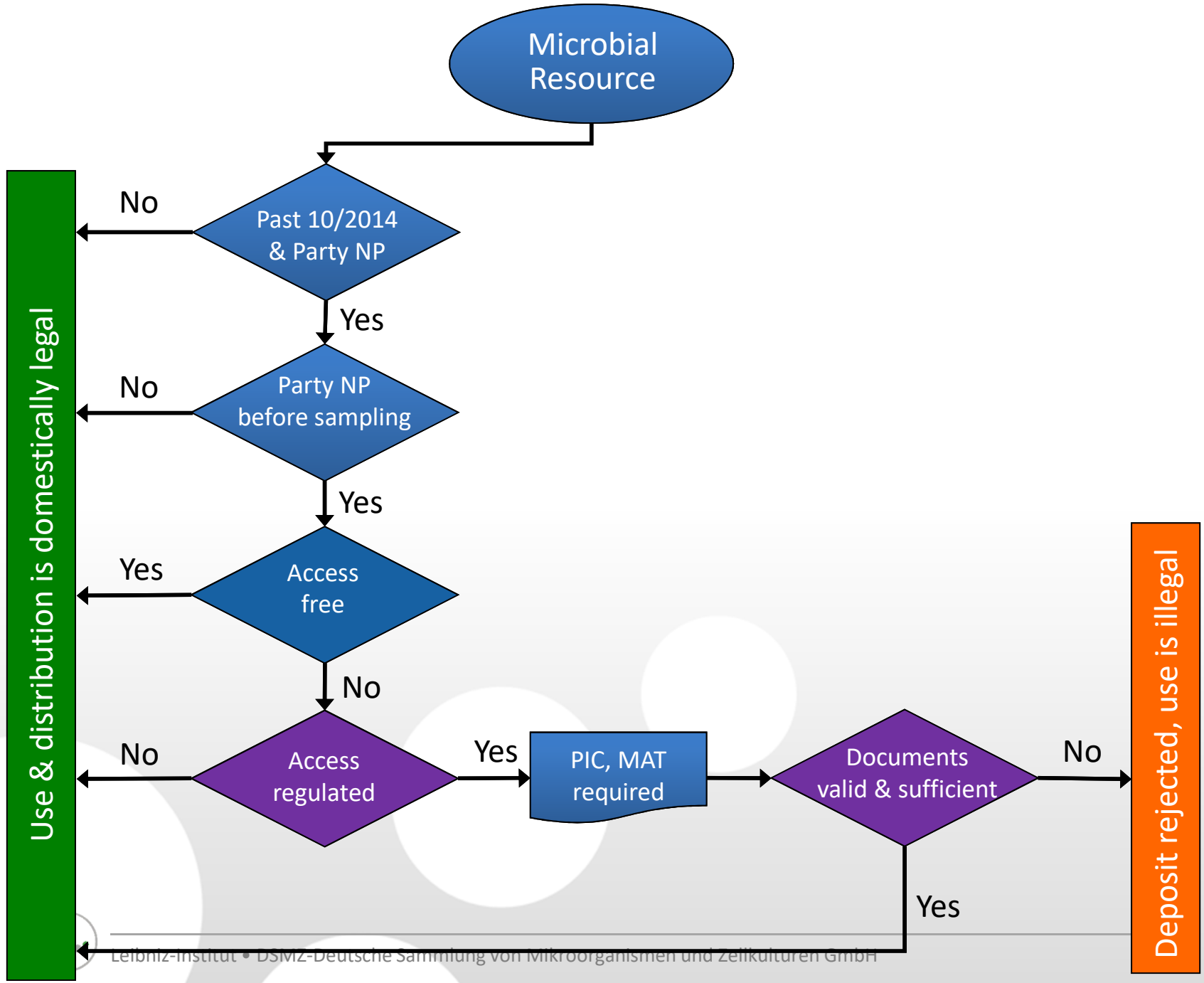
I have read and accept the DSMZ GmbH "Terms & Conditions" (or AGB in German, see **https://www.dsmz.de/terms.html**) and understand all of the obligations therein including:

If present, I will adhere to the requirements listed in the documents in the product's Nagoya protocol restrictions section of the online catalog and download and save the documents for 20 years after the last use of the product.

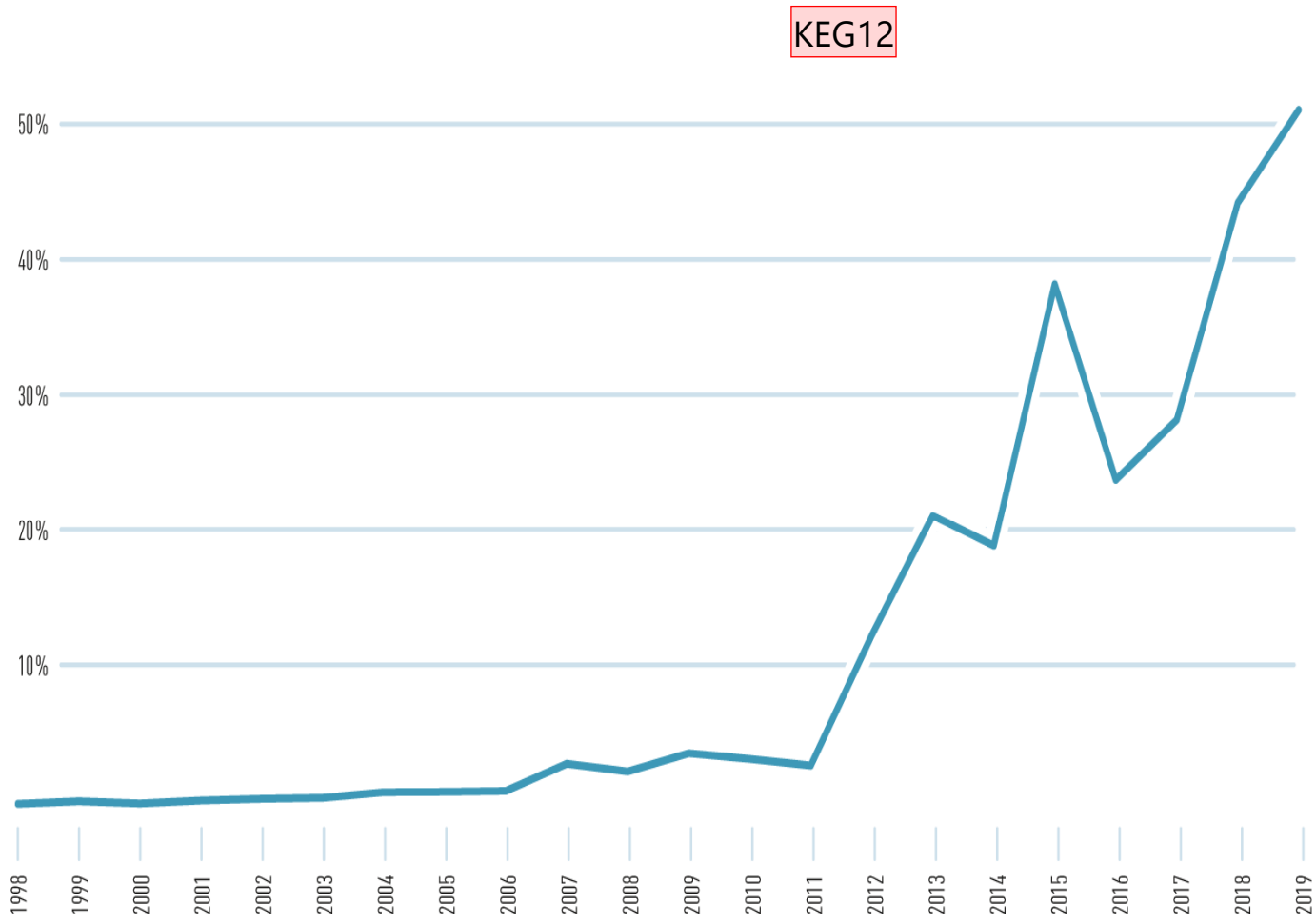
I will not distribute or share products with third parties or use products for commercial purposes. (Note: For some microbial strains, commercial use can be considered on a case-by-case basis. Please email sales@dsmz.de. For plant virus diagnostic purposes, commercial use is permitted. See "Terms & Conditions"/AGB.)

Questions?

amber.h.scholz@dsmz.de


Microbial Resource
Past 10/2014 & Party NP
No
Yes
Party NP before sampling
No
Yes
Access free
Yes
No
Access regulated
No
Yes
PIC, MAT required
Documents valid & sufficient
No
Yes
Use & distribution is domestically legal
Deposit rejected, use is illegal

How has the country tag changed over time?



Slide 18

KEG12 ??

Karger, Elizabeth GIZ; 25/10/2019

Entries vs. Information content