



“Open Data”

Open Access to Scientific Data

The perspective of a researcher

Prof. Dr. Tom Coenye

Tom.Coenye@UGent.be

Open Data – Shared Data

- Not necessarily the same !

- Data deposition (open/shared)



- Data available in Supporting Information Files (open/shared – but depending on OA status of publication)



- Data made available upon request (shared)

- Data available from third party (shared)

Data deposition

- Data can be deposited in an appropriate repository
 - Public repository
 - Other
 - Own repository/website
 - Institutional archive
 - ...
- Advantages of using public repositories over others
 - Continued access over time is guaranteed
 - An accession number, DOI, ... is assigned (~traceability)
 - No cost for researcher
 - Streamlined submission process
 - « Quality label »

Data deposition

- Data can deposited in an appropriate public repository
- Generalist repositories accept multiple data types
 - For example Dryad

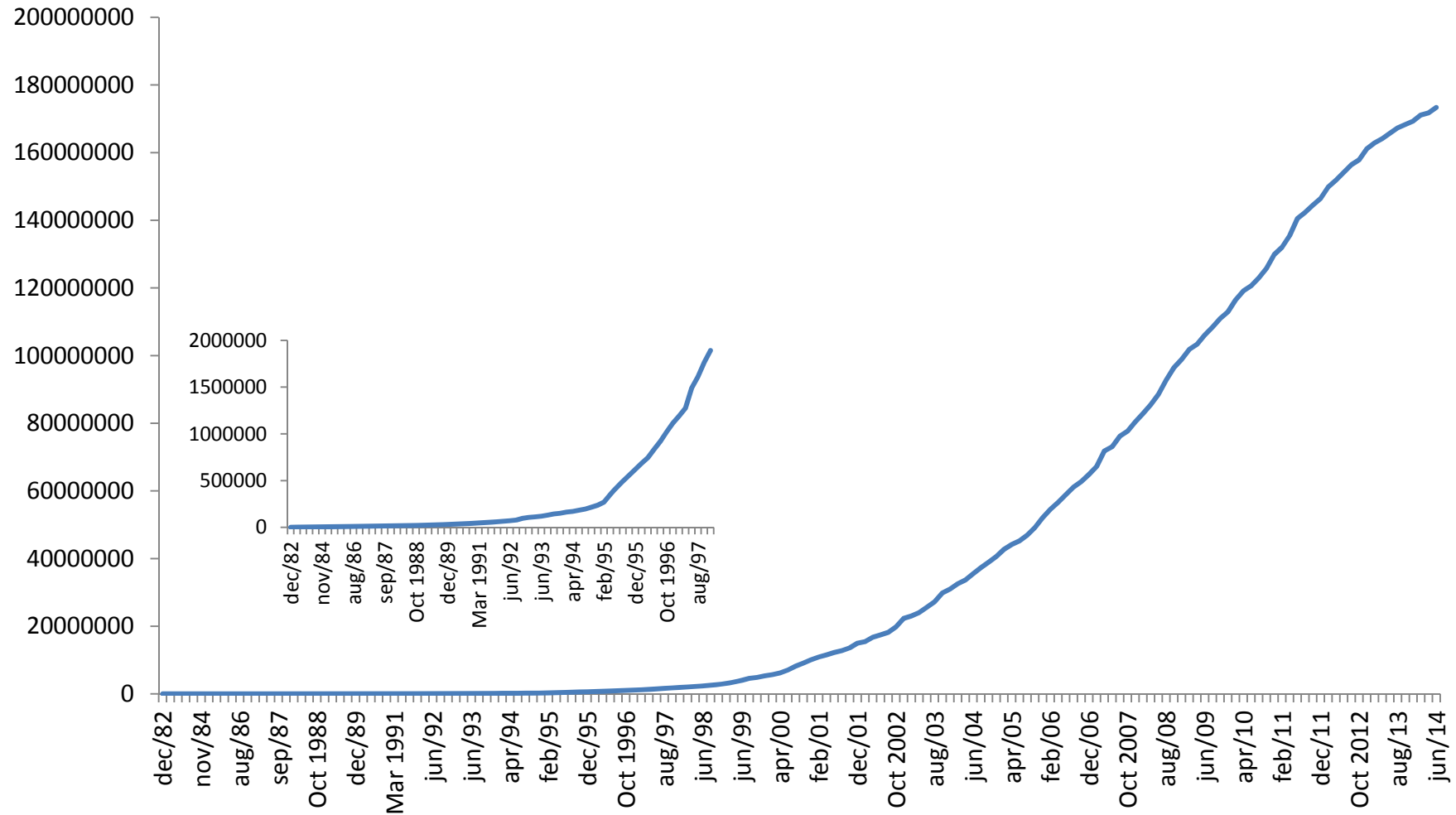


Data deposition

- Data can be deposited in an appropriate public repository
- Subject-specific repositories – examples from the Life Sciences
 - ArrayExpress or GEO – data on gene expression
 - GenBank, EMBL, DDBJ – nucleic acid sequence data
 - PDB – protein structures
 - ...

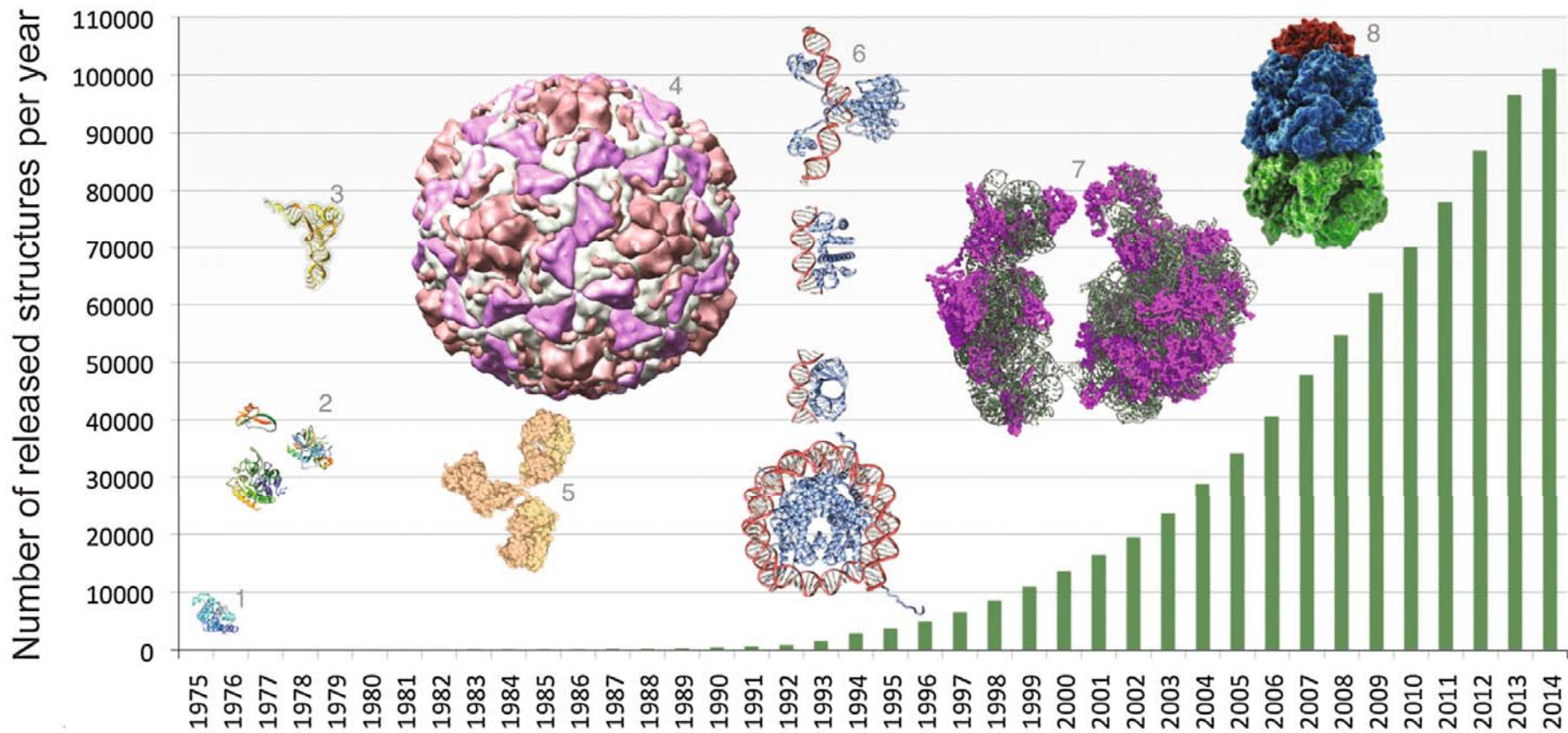
Public data deposition – nothing new!

•Size of GenBank database 1982-2014 (#entries)



Public data deposition – nothing new!

- Size of PDB Archive (1975-2014) (#entries)



Why would I share my data?

- Because I have to ...
 - Journal
 - University
 - Funding agency
 - ...
- Because it increases the visibility of my research



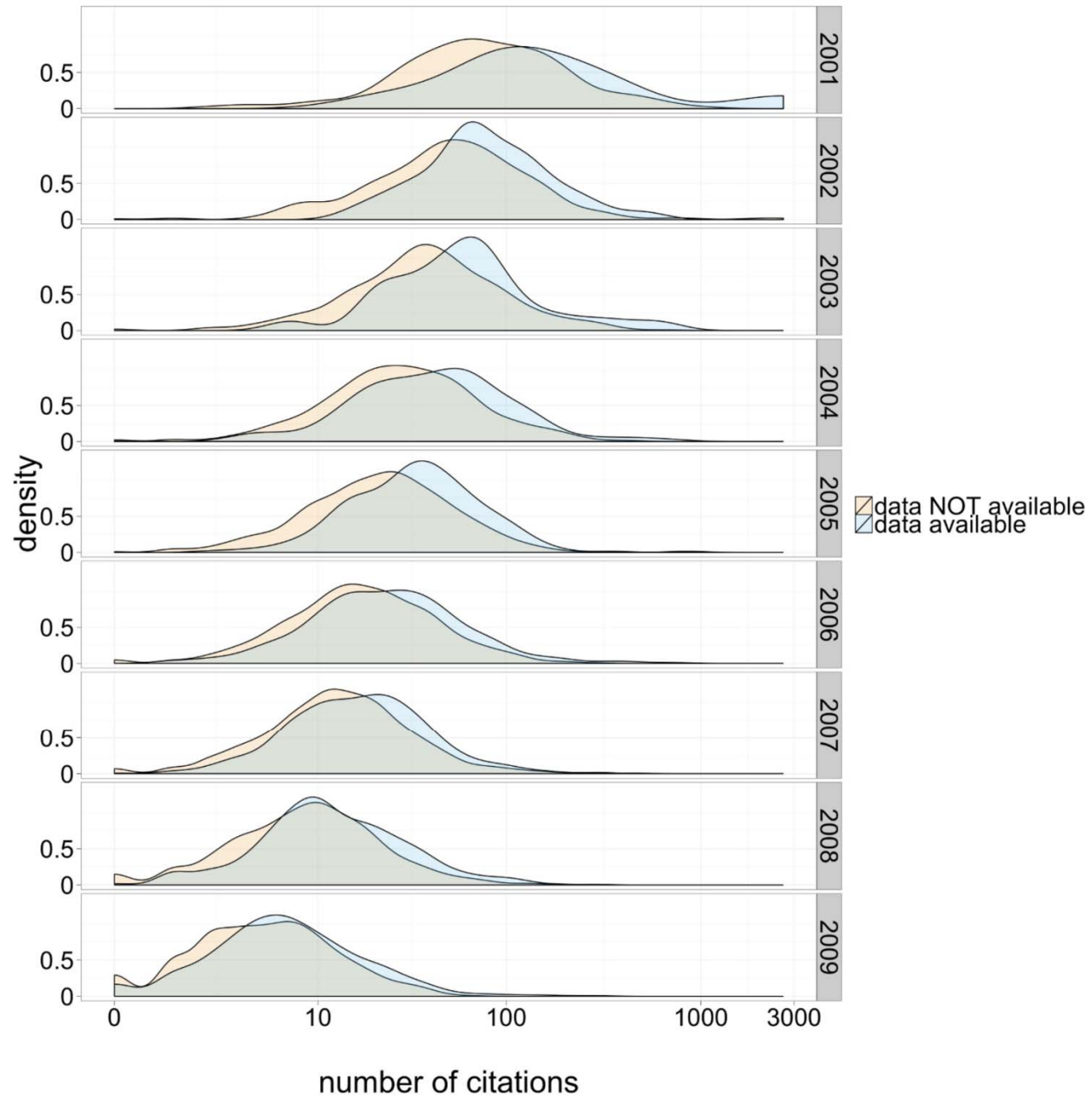
Data reuse and the open data citation advantage

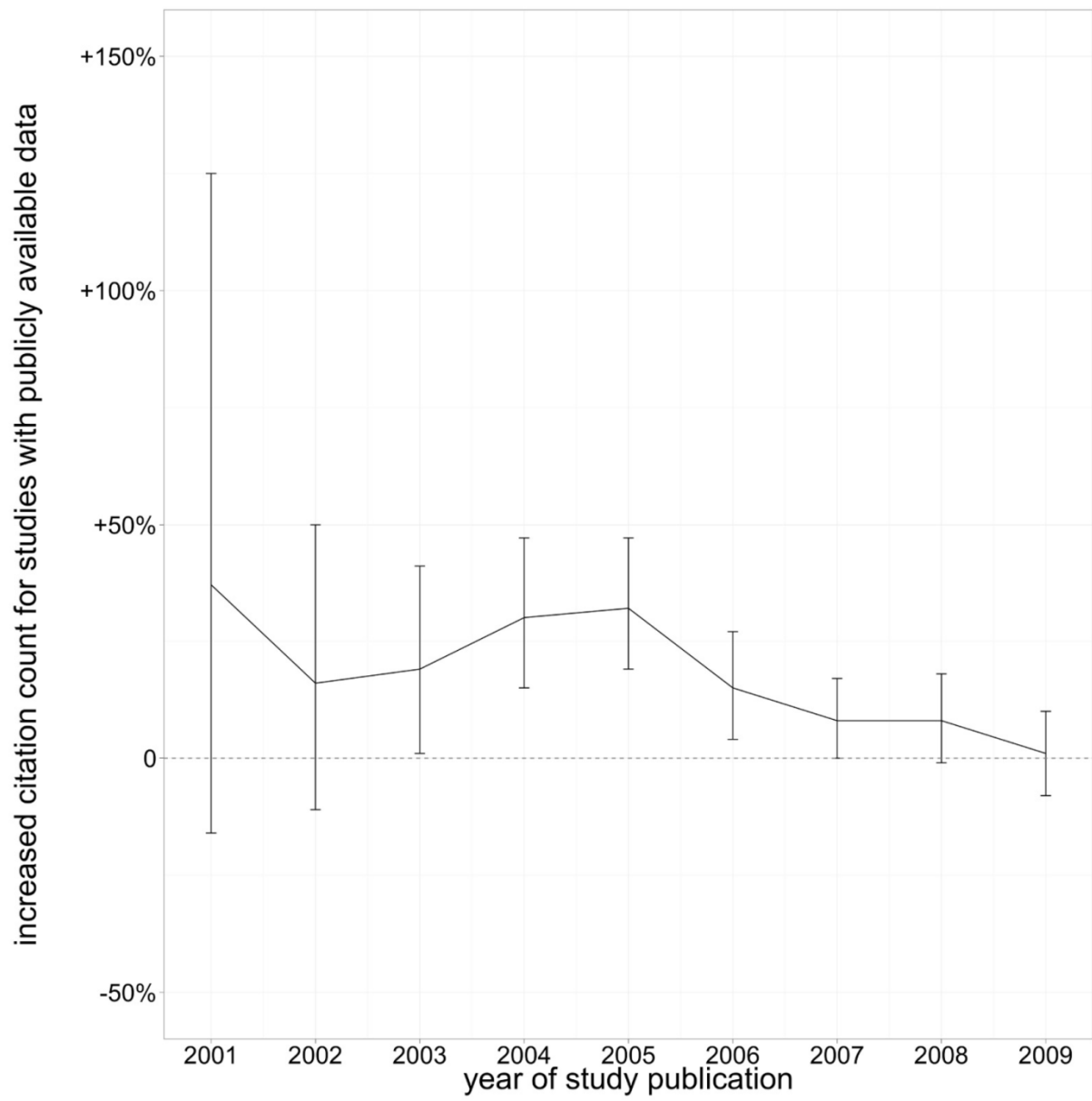
Heather A. Piwovar^{1,2} and Todd J. Vision^{1,2,3}

¹ National Evolutionary Synthesis Center, Durham, NC, USA

² Department of Biology, Duke University, Durham, NC, USA

³ Department of Biology, University of North Carolina - Chapel Hill, Chapel Hill, NC, USA





RESEARCH ARTICLE

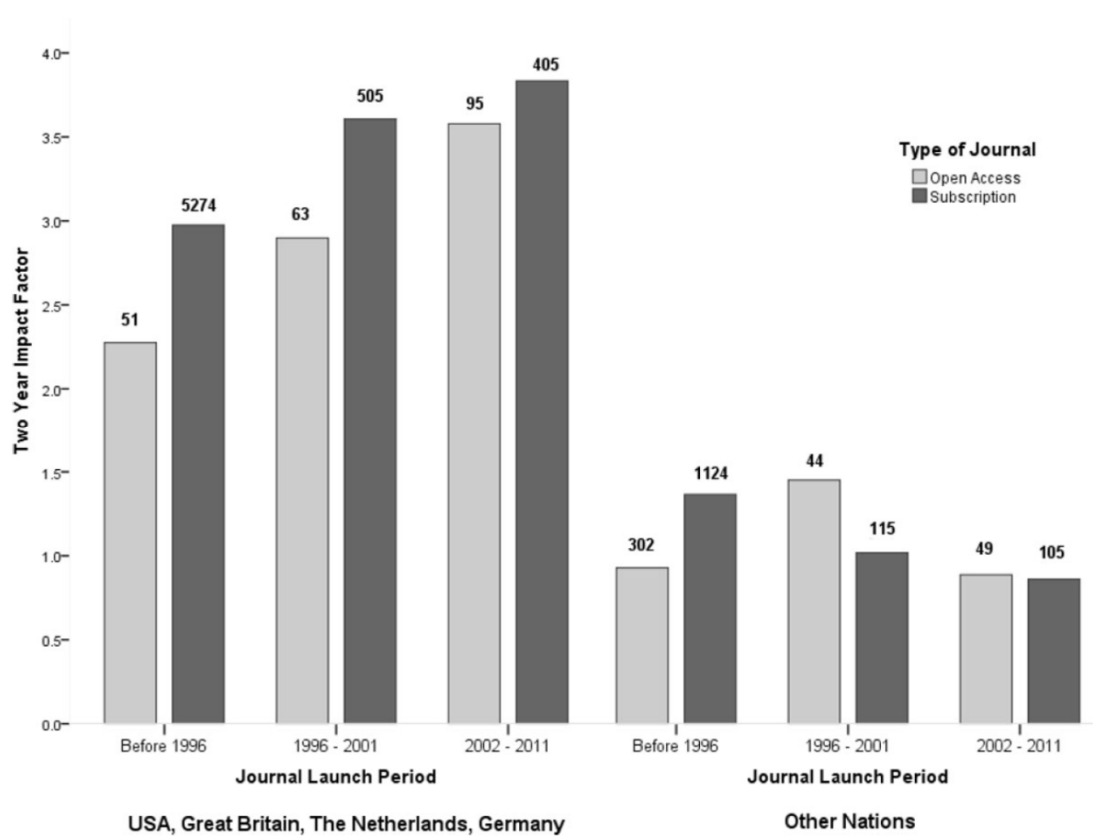
Open Access

Open access versus subscription journals:
a comparison of scientific impact

Bo-Christer Björk^{1*} and David Solomon²

- Differences between OA and subscription minimal at best

- Too early to tell....



Why would I share my data?

- Because it is good scientific practice and discourages QRP and scientific fraud
- Because it allows re-use of my data
 - By other scientists
 - By other stakeholders (policy makers, ...)
 - Brings data that were obtained with public funding in the public domain

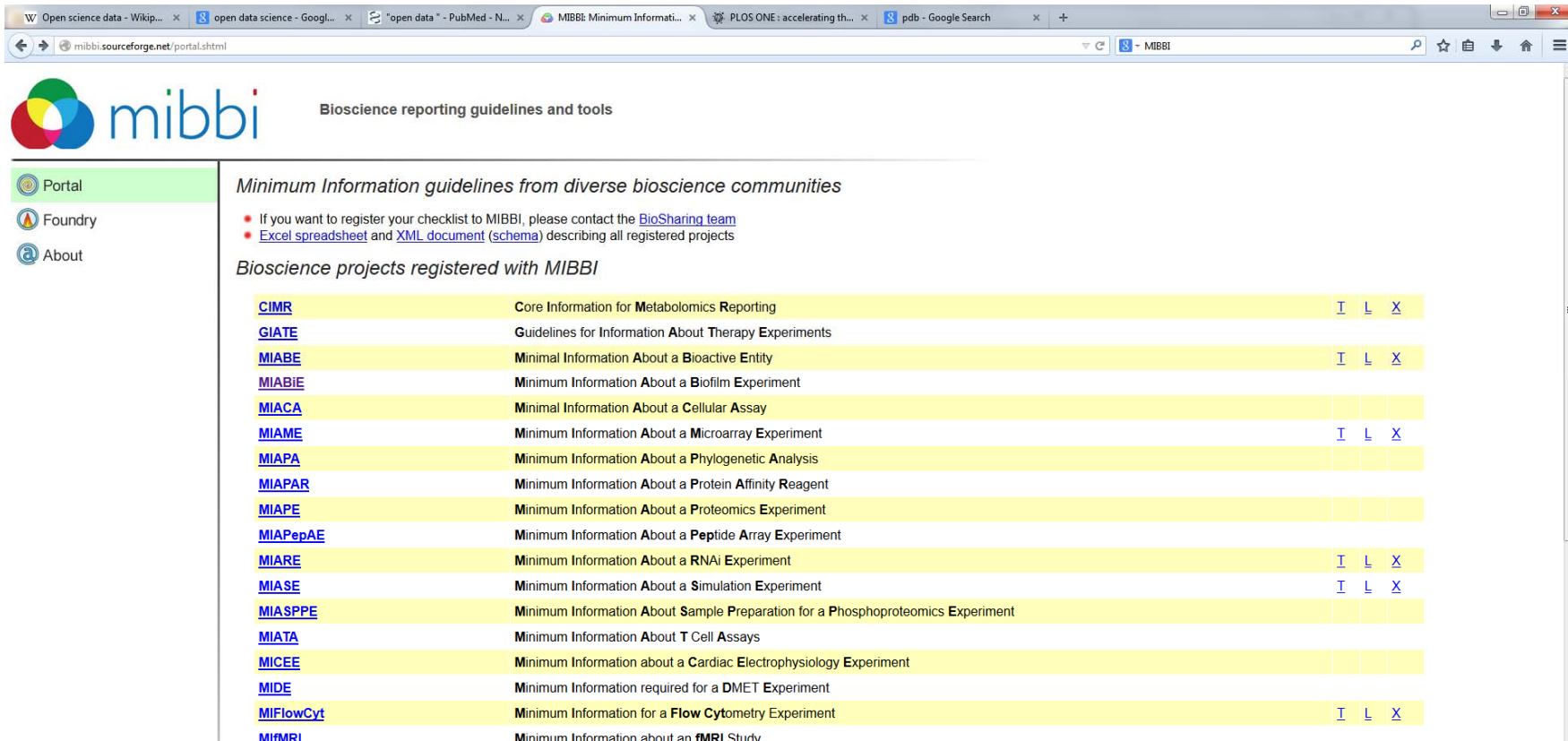
Why would I not share my data?

- Cost (*but as far as I know all public repositories in Life Sciences are free of charge for the researcher*)
- Time and effort (!!)
 - Overly complicated systems demanding a lot of time/effort will discourage deposition of data
 - Duplication of efforts should be avoided
- Competition – « I don't want others to be able to re-use my data! »
- Don't see added value – « What's in it for me? »

Public data deposition – a guarantee for quality or GIGO ?

Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project

<http://mibbi.sourceforge.net/portal.shtml>



The screenshot shows the MIBBI portal website. The header includes the MIBBI logo and the text "Bioscience reporting guidelines and tools". A left sidebar contains navigation links for "Portal", "Foundry", and "About". The main content area is titled "Minimum Information guidelines from diverse bioscience communities" and includes a list of registered projects with their respective guidelines and links to information, lists, and XML documents.

Minimum Information guidelines from diverse bioscience communities

- If you want to register your checklist to MIBBI, please contact the [BioSharing team](#)
- [Excel spreadsheet](#) and [XML document \(schema\)](#) describing all registered projects

Bioscience projects registered with MIBBI

CIMR	Core Information for Metabolomics Reporting	I	L	X
GIATE	Guidelines for Information About Therapy Experiments			
MIABE	Minimal Information About a Bioactive Entity	I	L	X
MIABIE	Minimum Information About a Biofilm Experiment			
MIACA	Minimal Information About a Cellular Assay			
MIAME	Minimum Information About a Microarray Experiment	I	L	X
MIAPA	Minimum Information About a Phylogenetic Analysis			
MIAPAR	Minimum Information About a Protein Affinity Reagent			
MIAPE	Minimum Information About a Proteomics Experiment			
MIAPepAE	Minimum Information About a Peptide Array Experiment			
MIARE	Minimum Information About a RNAi Experiment	I	L	X
MIASE	Minimum Information About a Simulation Experiment	I	L	X
MIASPPE	Minimum Information About Sample Preparation for a Phosphoproteomics Experiment			
MIATA	Minimum Information About T Cell Assays			
MICEE	Minimum Information about a Cardiac Electrophysiology Experiment			
MIDE	Minimum Information required for a DMET Experiment			
MIFlowCyt	Minimum Information for a Flow Cytometry Experiment	I	L	X
MIFMRI	Minimum Information about an fMRI Study			

The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments

Stephen A. Bustin,^{1*} Vladimir Benes,² Jeremy A. Garson,^{3,4} Jan Hellemans,⁵ Jim Hugge
Mikael Kubista,^{7,8} Reinhold Mueller,⁹ Tania Nolan,¹⁰ Michael W. Pfaffl,¹¹ Gregory L. Ship
Jo Vandesompele,⁵ and Carl T. Wittwer^{13,14}

<http://www.rdml.org/miqe.php>

Table 1. MIQE checklist for authors, reviewers, and editors.^a

Item to check	Importance	Item to check	Importance
Experimental design		qPCR oligonucleotides	
Definition of experimental and control groups	E	Primer sequences	E
Number within each group	E	RTPrimerDB identification number	D
Assay carried out by the core or investigator's laboratory?	D	Probe sequences	D ^d
Acknowledgment of authors' contributions	D	Location and identity of any modifications	E
Sample		Manufacturer of oligonucleotides	
Description	E	Purification method	D
Volume/mass of sample processed	D	qPCR protocol	
Microdissection or macrodissection	E	Complete reaction conditions	E
Processing procedure	E	Reaction volume and amount of cDNA/DNA	E
If frozen, how and how quickly?	E	Primer, (probe), Mg ²⁺ , and dNTP concentrations	E
If fixed, with what and how quickly?	E	Polymerase identity and concentration	E
Sample storage conditions and duration (especially for FFPE ^b samples)	E	Buffer/kit identity and manufacturer	E
Nucleic acid extraction		Exact chemical composition of the buffer	
Procedure and/or instrumentation	E	Additives (SYBR Green I, DMSO, and so forth)	E
Name of kit and details of any modifications	E	Manufacturer of plates/tubes and catalog number	D
Source of additional reagents used	D	Complete thermocycling parameters	E
Details of DNase or RNase treatment	E	Reaction setup (manual/robotic)	D
Contamination assessment (DNA or RNA)	E	Manufacturer of qPCR instrument	E
Nucleic acid quantification		qPCR validation	
Instrument and method	E	Evidence of optimization (from gradients)	D
Purity (A ₂₆₀ /A ₂₈₀)	D	Specificity (gel, sequence, melt, or digest)	E
Yield	D	For SYBR Green I, C _q of the NTC	E
RNA integrity: method/instrument	E	Calibration curves with slope and y intercept	E
RIN/RQI or C _q of 3' and 5' transcripts	E	PCR efficiency calculated from slope	E
Electrophoresis traces	D	CIs for PCR efficiency or SE	D
Inhibition testing (C _q dilutions, spike, or other)	E	r ² of calibration curve	E
Reverse transcription		Linear dynamic range	
Complete reaction conditions	E	C _q variation at LOD	E
Amount of RNA and reaction volume	E	CIs throughout range	D
Priming oligonucleotide (if using GSP) and concentration	E	Evidence for LOD	E
Reverse transcriptase and concentration	E	If multiplex, efficiency and LOD of each assay	E
Temperature and time		Data analysis	
Manufacturer of reagents and catalogue numbers	D	qPCR analysis program (source, version)	E
C _q with and without reverse transcription	D ^c	Method of C _q determination	E
Storage conditions of cDNA	D	Outlier identification and disposition	E
qPCR target information		Results for NTCs	
Gene symbol	E	Justification of number and choice of reference genes	E
Sequence accession number	E	Description of normalization method	E
Location of amplicon	D	Number and concordance of biological replicates	D
Amplicon length	E	Number and stage (reverse transcription or qPCR) of technical replicates	E
In silico specificity screen (BLAST, and so on)	E	Repeatability (intraassay variation)	E
Pseudogenes, retropseudogenes, or other homologs?	D	Reproducibility (interassay variation, CV)	D
Sequence alignment	D	Power analysis	D
Secondary structure analysis of amplicon	D	Statistical methods for results significance	E
Location of each primer by exon or intron (if applicable)	E	Software (source, version)	E
What splice variants are targeted?	E	C _q or raw data submission with RDML	D

^aAll essential information (E) must be submitted with the manuscript. Desirable information (D) should be submitted if available. If primers are from RTPrimerDB, information on qPCR target, oligonucleotides, protocols, and validation is available from that source.

^bFFPE, formalin-fixed, paraffin-embedded; RIN, RNA integrity number; RQI, RNA quality indicator; GSP, gene-specific priming; dNTP, deoxynucleoside triphosphate.

^cAssessing the absence of DNA with a no-reverse transcription assay is essential when first extracting RNA. Once the sample has been validated as rDNA free, inclusion of a no-reverse transcription control is desirable but no longer essential.

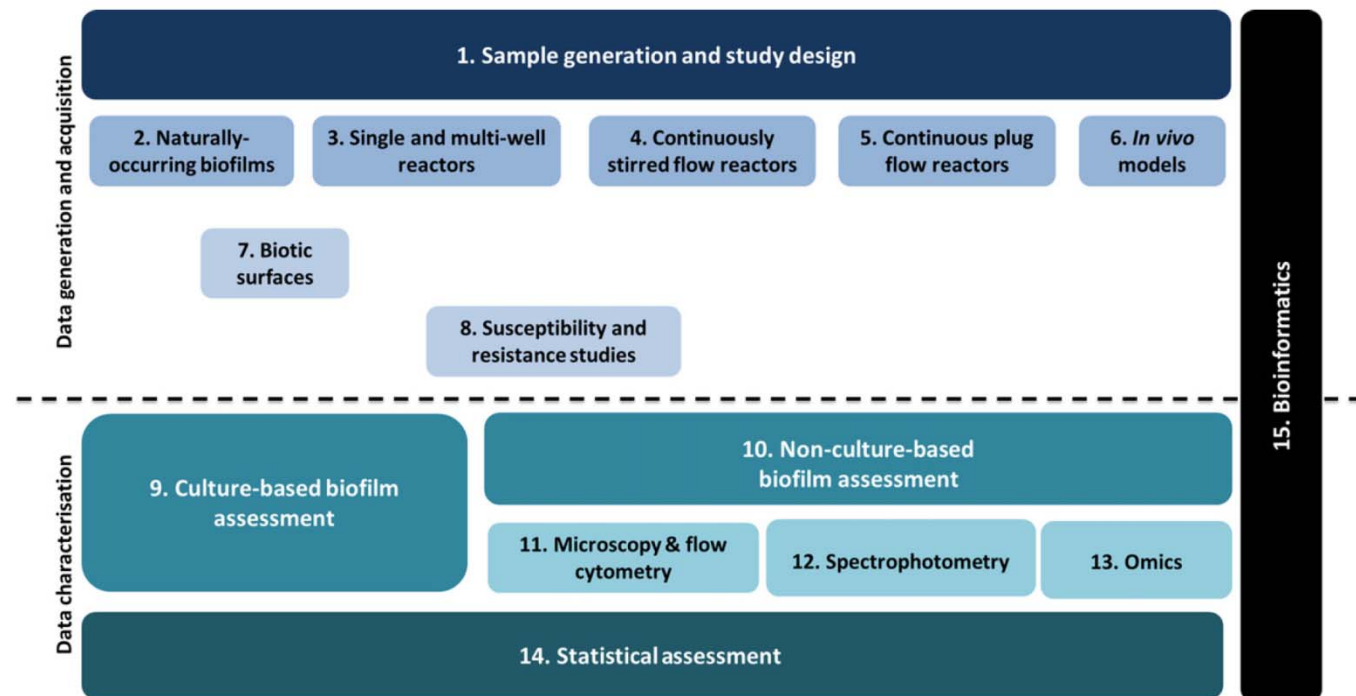
^dDisclosure of the probe sequence is highly desirable and strongly encouraged; however, because not all vendors of commercial predesigned assays provide this information, it cannot be an essential requirement. Use of such assays is discouraged.

MINIREVIEW

Minimum information about a biofilm experiment (MIABiE): standards for reporting experiments and data on sessile microbial communities living at interfaces

Anália Lourenço^{1,2}, Tom Coenye³, Darla M. Goeres⁴, Gianfranco Donelli⁵, Andreia S. Azevedo⁶, Howard Ceri⁷, Filipa L. Coelho², Hans-Curt Flemming⁸, Talis Juhna^{9,10}, Susana P. Lopes², Rosário Oliveira², Antonio Oliver¹¹, Mark E. Shirtliff^{12,13}, Ana M. Sousa², Paul Stoodley¹⁴, Maria Olivia Pereira² & Nuno F. Azevedo⁶

<http://biofomics.org/>



Control Control Control: A Reassessment and Comparison of GenBank and Chromatogram mtDNA Sequence Variation in Baltic Grey Seals (*Halichoerus grypus*)

Katharina Fietz^{1*}, Jeff A. Graves², Morten Tange Olsen^{1,3*}

“By re-editing the original chromatogram data we found that **approximately 40% of the grey seal mtDNA haplotype sequences posted in GenBank contained errors**. ... a **significantly different outcome was observed** when using the uncorrected dataset based on the GenBank haplotypes. We therefore suggest disregarding the existing GenBank data and instead using the correct haplotypes reported here.”

“Our study serves as an illustrative example reiterating the importance of quality control through every step of a research project, from data generation to interpretation and submission to an online repository. Errors conducted in any step may lead to biased results and conclusions, and could impact management decisions.”

What needs to be deposited and how?

- ALL data ?
 - Raw data?
 - Processed data?
 - Selection of raw and/or processed data?

- Minimal dataset ?
 - PLoS One : *the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety*

- What about biological material ?
 - Cell lines, tissues, animals, purified biomolecules, ...
 - Mandatory deposition in BRCs ?
 - Similar issues likely with chemicals etc.



“Open Data”

Open Access to Scientific Data

The perspective of a researcher

Prof. Dr. Tom Coenye

Tom.Coenye@UGent.be