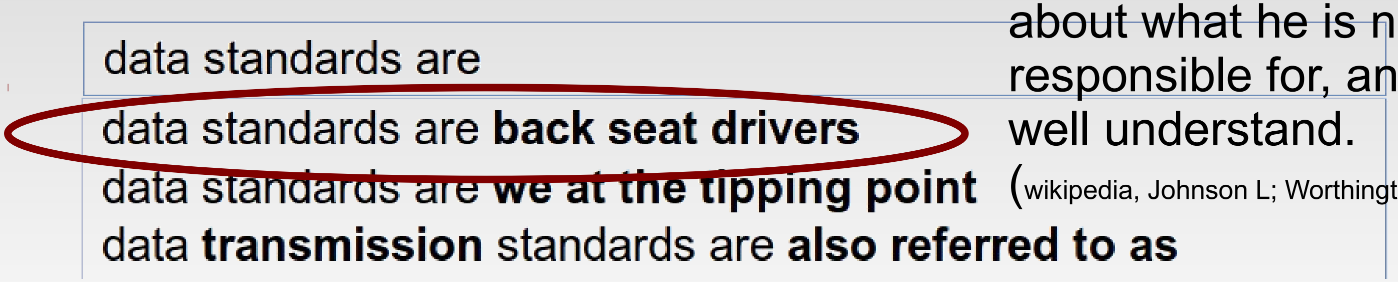
Research data standards and -sharing in Lifesciences

Steffen Neumann
Leibniz Institute of Plant Biochemistry
Mass Spectrometry and Bioinformatics group
sneumann@ipb-halle.de

Google



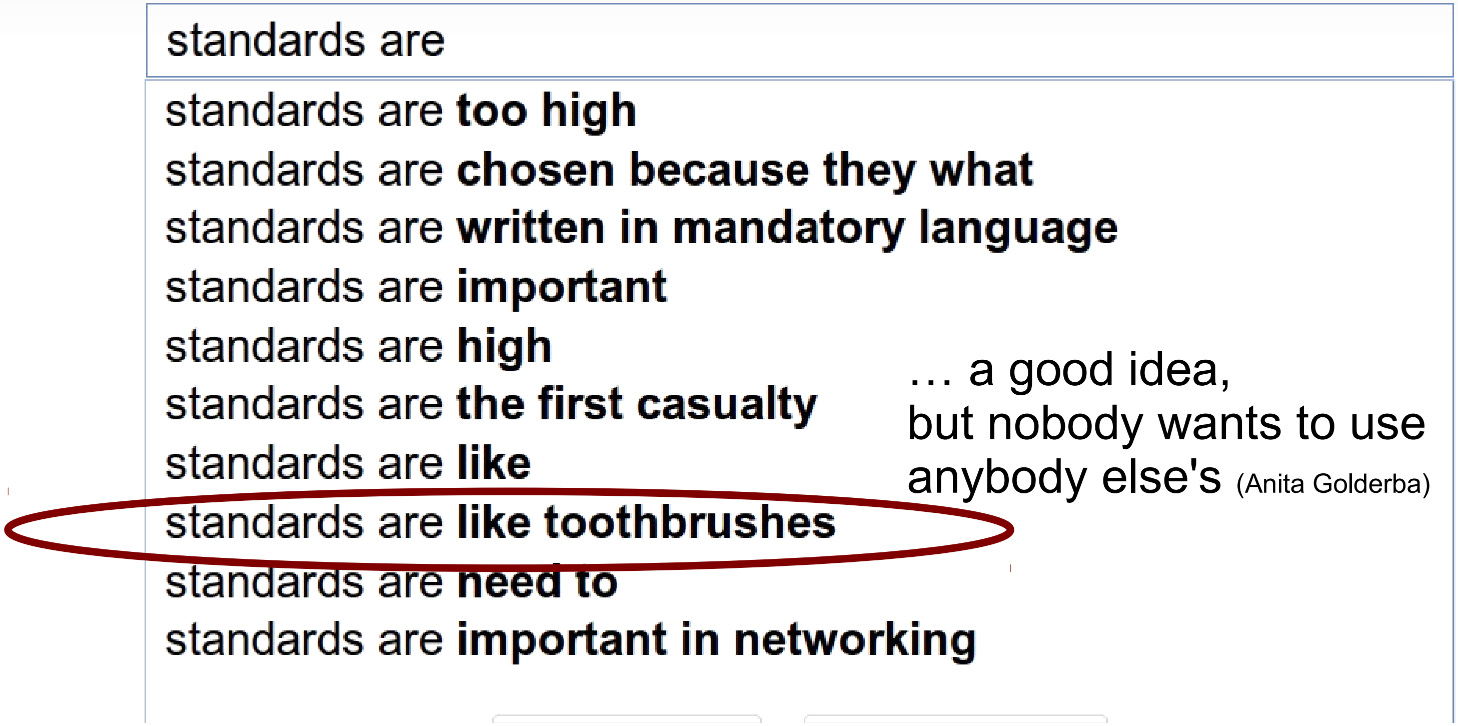
data standards are

data standards are **back seat drivers**

data standards are **we at the tipping point**

data **transmission** standards are **also referred to as**

a person who gives advice about what he is not responsible for, and may not well understand. (wikipedia, Johnson L; Worthington J.C)



standards are

standards are **too high**

standards are **chosen because they what**

standards are **written in mandatory language**

standards are **important**

standards are **high**

standards are **the first casualty**

standards are **like**

standards are **like toothbrushes**

standards are **need to**

standards are **important in networking**

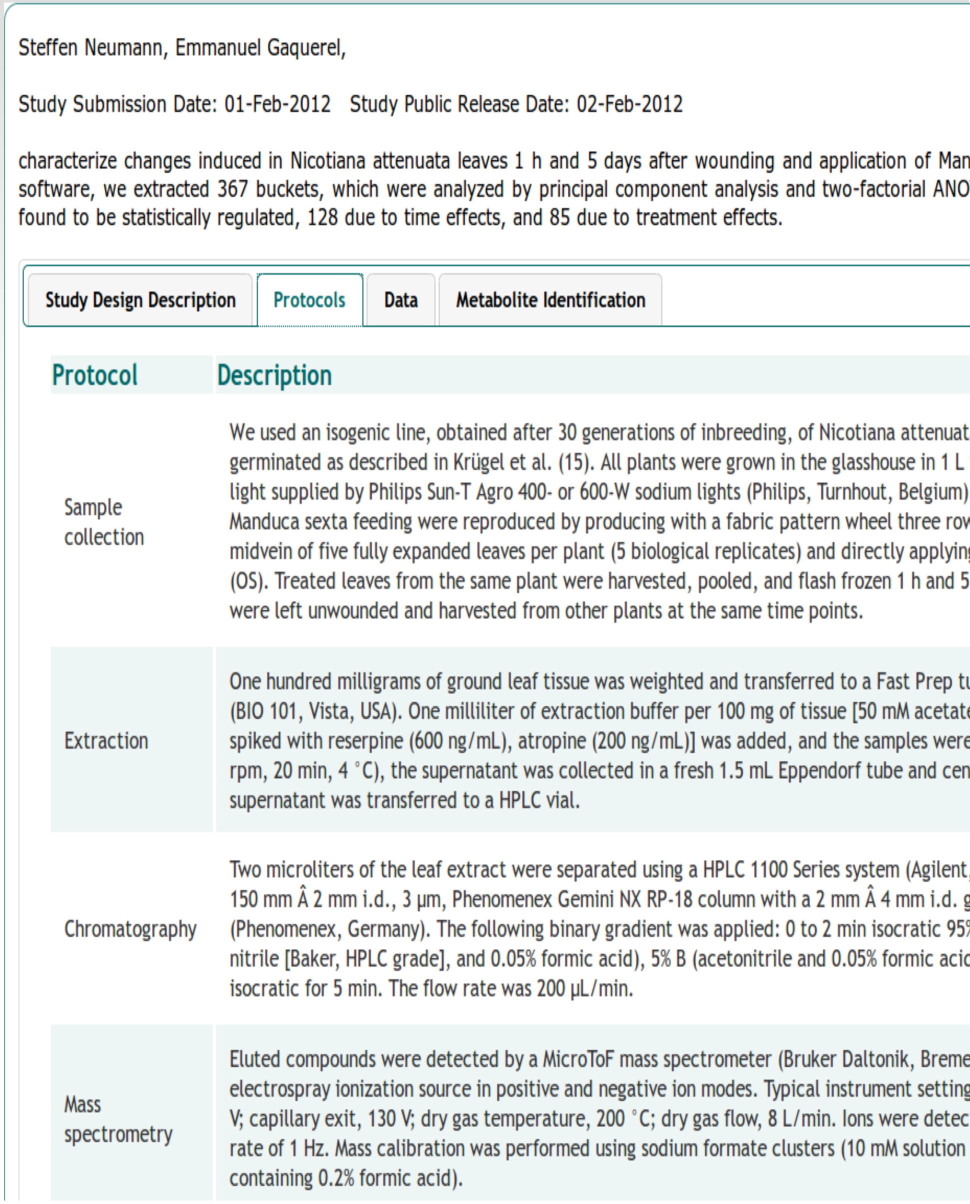
… a good idea, but nobody wants to use anybody else's (Anita Golderba)

How do we report Metabolomics results ?

- We write papers which have "Materials and Methods" and "Results" sections
- The gene expression data community was among the first to draft reporting guidelines "MIAME" Minimum Information About Microarray Experiment
- The metabolomics community also had efforts to suggest what to report:
  - Standardised Metabolomics Reporting Structure (SMRS)
  - Metabolomics Standards Initiative (MSI)
- There were databases and schema definitions (ArMet, CIMR)

MetaboLights@EBI

- Repository of Metabolomics Data
- ISAtab metadata format
- Array Express, Pride, GenBank, …
- IPB Halle investigates use as in-house research data repository

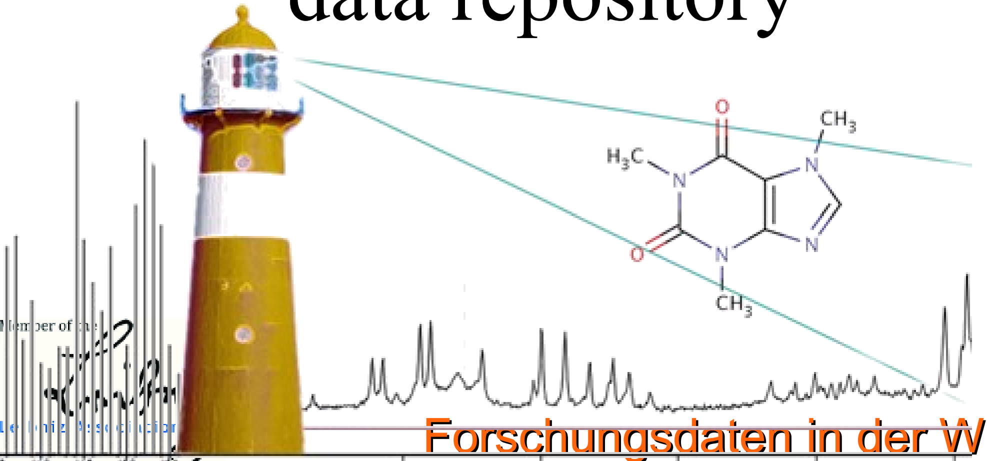


Steffen Neumann, Emmanuel Gaquerel,

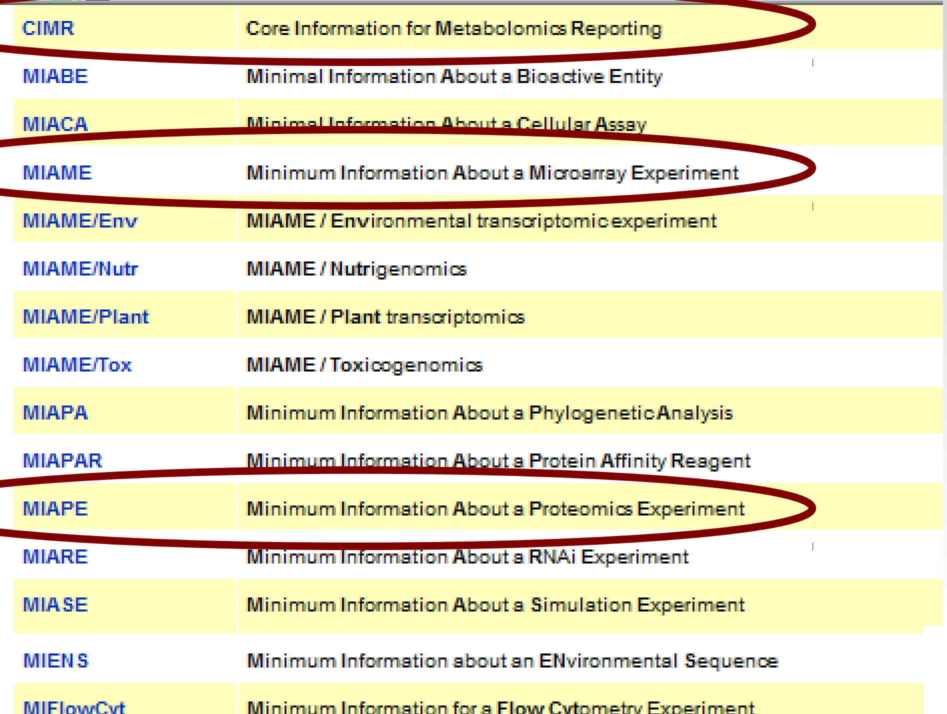
Study Submission Date: 01-Feb-2012   Study Public Release Date: 02-Feb-2012

characterize changes induced in Nicotiana attenuata leaves 1 h and 5 days after wounding and application of Man
software, we extracted 367 buckets, which were analyzed by principal component analysis and two-factorial ANO
found to be statistically regulated, 128 due to time effects, and 85 due to treatment effects.

Protocol	Description
Sample collection	We used an isogenic line, obtained after 30 generations of inbreeding, of Nicotiana attenuat germinated as described in Krügel et al. (15). All plants were grown in the glasshouse in 1 L light supplied by Philips Sun-T Agro 400- or 600-W sodium lights (Philips, Turnhout, Belgium) Manduca sexta feeding were reproduced by producing with a fabric pattern wheel three row midvein of five fully expanded leaves per plant (5 biological replicates) and directly applyin (OS). Treated leaves from the same plant were harvested, pooled, and flash frozen 1 h and 5 were left unwounded and harvested from other plants at the same time points.
Extraction	One hundred milligrams of ground leaf tissue was weighted and transferred to a Fast Prep tu (BIO 101, Vista, USA). One milliliter of extraction buffer per 100 mg of tissue [50 mM acetat spiked with reserpine (600 ng/mL), atropine (200 ng/mL)] was added, and the samples were rpm, 20 min, 4 °C), the supernatant was collected in a fresh 1.5 mL Eppendorf tube and cen supernatant was transferred to a HPLC vial.
Chromatography	Two microliters of the leaf extract were separated using a HPLC 1100 Series system (Agilent, 150 mm Â 2 mm i.d., 3 μm, Phenomenex Gemini NX RP-18 column with a 2 mm Â 4 mm i.d. g (Phenomenex, Germany). The following binary gradient was applied: 0 to 2 min isocratic 95% nitrile [Baker, HPLC grade], and 0.05% formic acid), 5% B (acetonitrile and 0.05% formic acid isocratic for 5 min. The flow rate was 200 μL/min.
Mass spectrometry	Eluted compounds were detected by a MicroToF mass spectrometer (Bruker Daltonik, Breme electrospray ionization source in positive and negative ion modes. Typical instrument setting V; capillary exit, 130 V; dry gas temperature, 200 °C; dry gas flow, 8 L/min. Ions were detec rate of 1 Hz. Mass calibration was performed using sodium formate clusters (10 mM solution containing 0.2% formic acid).

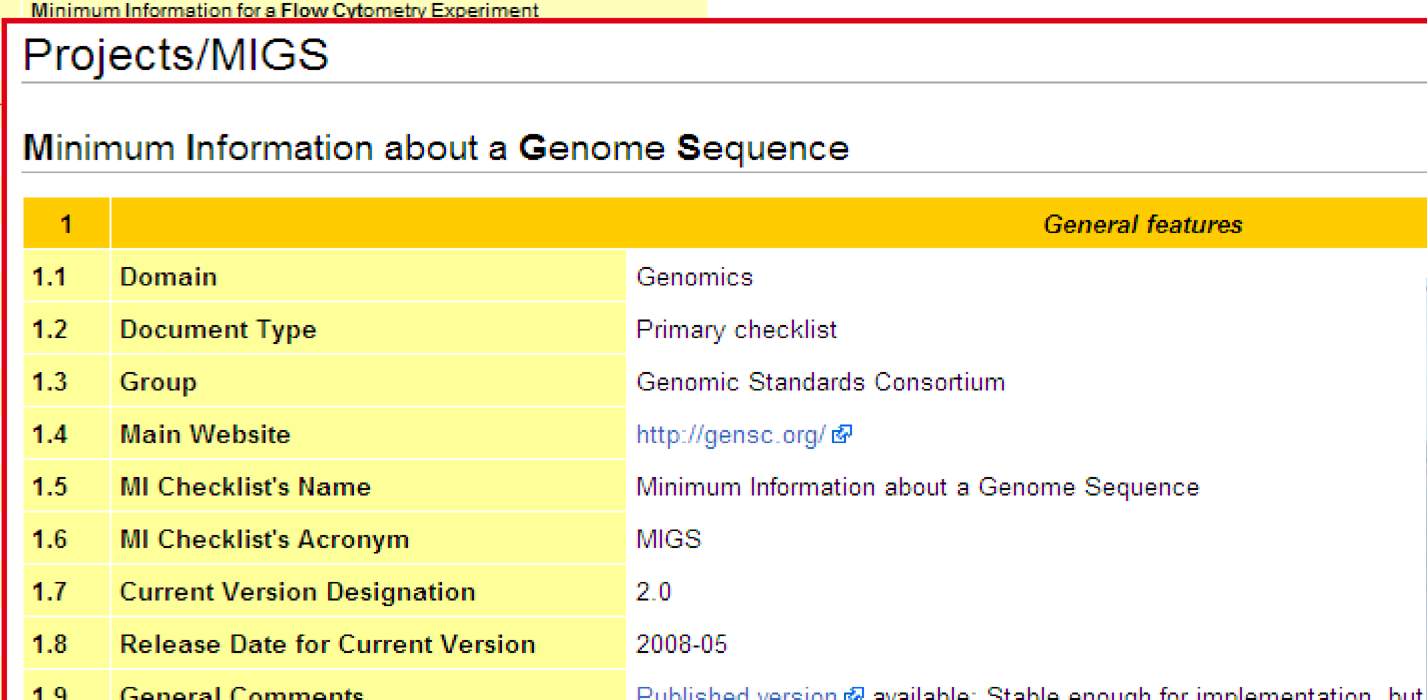


What about other -omics ?



CIMR	Core Information for Metabolomics Reporting
MIABE	Minimal Information About a Bioactive Entity
MIACA	Minimal Information About a Cellular Assay
MIAME	Minimum Information About a Microarray Experiment
MIAME/Env	MIAME / Environmental transcriptomic experiment
MIAME/Nutr	MIAME / Nutrigenomics
MIAME/Plant	MIAME / Plant transcriptomics
MIAME/Tox	MIAME / Toxicogenomics
MIAPA	Minimum Information About a Phylogenetic Analysis
MIAPAR	Minimum Information About a Protein Affinity Reagent
MIAPE	Minimum Information About a Proteomics Experiment
MIARE	Minimum Information About a RNAi Experiment
MIASE	Minimum Information About a Simulation Experiment
MIENS	Minimum Information about an ENvironmental Sequence
MIFlowCyt	Minimum Information for a Flow Cytometry Experiment
MIGen	
MIGS	
MIMIx	
MIMPP	
MINI	
MINIMESS	
MINSEQE	
MIPFE	
MIQAS	
MIqPCR	
MIRIAM	
MISFISHIE	
STRENDA	
TBC	

- www.mibbi.org
- Minimum Information for Biological and Biomedical Investigations
- Reporting guidelines from *many* disciplines in life sciences



Projects/MIGS

Minimum Information about a Genome Sequence

1	General features	
1.1	Domain	Genomics
1.2	Document Type	Primary checklist
1.3	Group	Genomic Standards Consortium
1.4	Main Website	http://gensc.org/
1.5	MI Checklist's Name	Minimum Information about a Genome Sequence
1.6	MI Checklist's Acronym	MIGS
1.7	Current Version Designation	2.0
1.8	Release Date for Current Version	2008-05
1.9	General Comments	Published version available; Stable enough for implementation, but

The MIBBI Project (mibbi.org)

Comparison of MIBBI-registered projects [21]

Version 0.7 (2008-04-10)

Granularity	🟩 Coarse	🟨 Medium	🟥 Fine
Maturity	⚫ Planned	🔵 Drafting	🟢 Release

[†] Denotes that a specification is provided as a suite of related doc

CONCEPT	SPECIALISATION	CIMR [†] 🔴	MIACA 🟢	MIAME 🔴	MIAME/Env 🟢	MIAME/Nutr 🟢	MIAME/Plant 🔴	MIAME/Tox 🟢	MIAPA ⚫	MIAPE [†] 🔴	MIARE 🟢	MIFlowCyt 🔴	MIGen 🔵	MIGS/MIMS 🔴	MIMIx 🔴	MIMPP 🔵	MINI 🔵
study inputs	study design	🟥	🟨	🟩	🟨	🟨	🟨	🟩	🟨	🟨 ⚫	🟨	🟩	🟨			🟩	🟩
	generic organism	🟥	🟩	🟨	🟨	🟥		🟥	🟩	🟨 ⚫	🟨	🟨	🟨	🟨	🟩	🟨	🟩
	cells / microbes	🟥	🟥			🟥		🟨				🟩	🟩				🟩
	plant	🟥					🟥										
	animal	🟥				🟥		🟨									
	mouse															🟥	
	human	🟨											🟨				🟩
	population	🟩											🟥				
	environmental sample				🟨									🟨			
	environment / habitat	🟥			🟥		🟨					🟩	🟩	🟨			
	in silico model								🟨								
study procedures	organism maintenance				🟨												
	animal husbandry	🟥				🟨		🟨								🟩	
	cell / microbe culture	🟥	🟥			🟨		🟨									
	plant cultivation	🟥					🟥										
	acclimation	🟨															
	preconditioning / pretreatment	🟥			🟥		🟥			🟨 ⚫	🟨						🟩
	organism manipulation	🟨				🟥		🟨					🟨				
assay inputs	generic study input			🟨	🟨	🟨	🟨	🟨									
	organism part	🟥	🟩	🟩	🟩	🟨	🟩	🟩		🟨 ⚫	🟩	🟩	🟨				🟨
	organism state	🟨					🟨					🟩				🟥	🟨
	organism trait											🟨	🟥	🟨		🟥	
	biomolecule		🟩												🟨		
	synthetic analyte		🟨		🟩		🟩			🟨 ⚫							

Granularity of metadata

- There can be too little and too much metadata
  - Too little: meaningless data dump
  - Too much: over-engineered database, problematic User Acceptance
- Future: Text-Mining the MetaData ?
  - Automatic RDF extraction as suggestion during submission
- Different granularity for in-house “LIMS” and published data-sets ?

License: rights+obligations

- CreativeCommons: Family of licenses
  - Zero (CC0): Data fully in the public domain
  - ByAttribution (BY): Obligation to cite the origin
  - ShareAlike (SA): Derivatives must be CC-SA as well
    The “viral” license, it spreads!
  - NoDerivatives (ND): Redistribution only unchanged
  - NonCommercial (NC): problematic definition
  - And (almost) combinations thereof
- And many other: `opendatacommons.org/licenses`
- **No license is the worst choice**, because the “default rights/obligations” differ across legislations

Member of the Leibniz Association

Take Home

- There are definitions for Metadata out there
  - Lots of them on mibbi.org
- ISA tools can capture them
  - In-house and public MetaboLights
- What's left is to **agree** and use all of the above
  - Ongoing EU FP7 Coordinating action: “COSMOS” COordination Of Standards in Metabolomics

Summary

- Data Sharing is a technical and political challenge
- (Meta-) Data Standards must help,
  not impede adoption by consumers and producers
- There are both successful and unsuccessful examples to learn from
- Quote: “Lead-by-example”