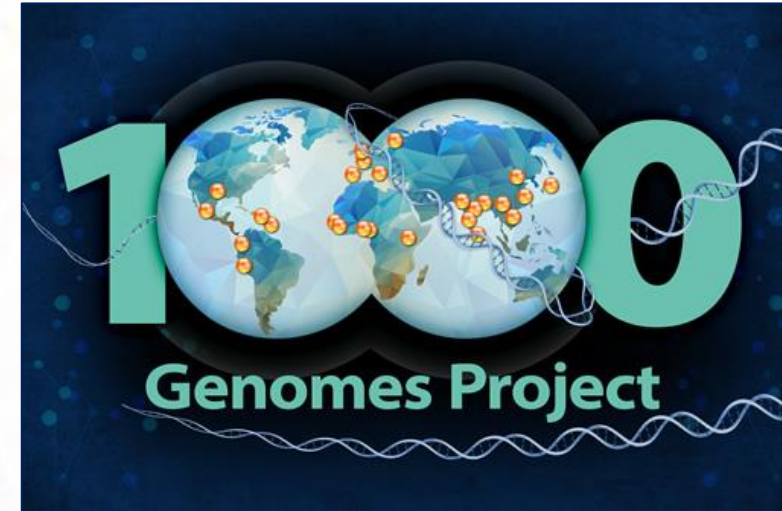
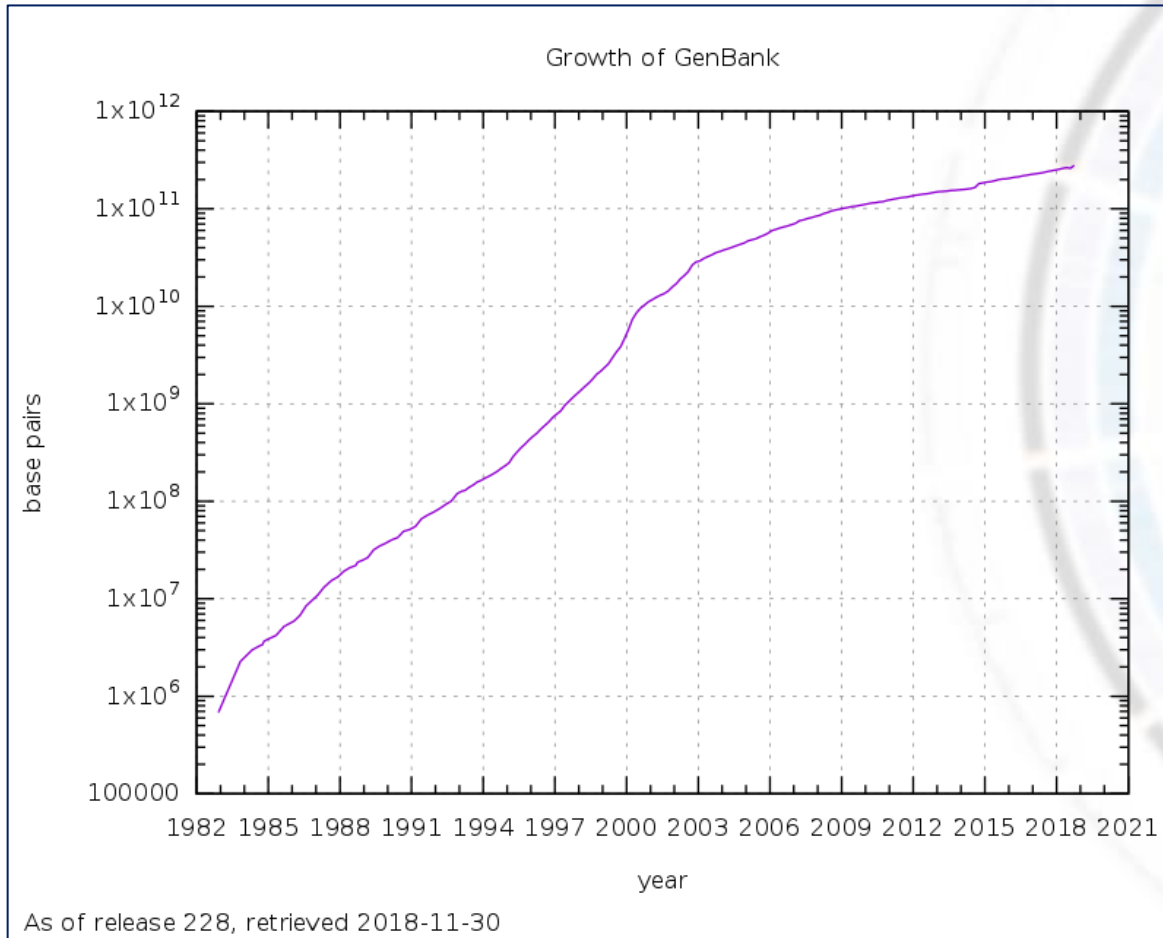



Noções gerais de trabalho em ambiente Linux e em linha de comando

Dr^a Desirrê Petters-Vandresen

Grande volume de dados disponíveis, em constante crescimento...



JGI  **MycoCosm**
THE FUNGAL GENOMICS RESOURCE

[Home](#) [Outreach](#) [Video Tutorials](#) [About](#)

1000 Fungal Genomes Project

[Nominate a genome to sequence](#)

Desafios

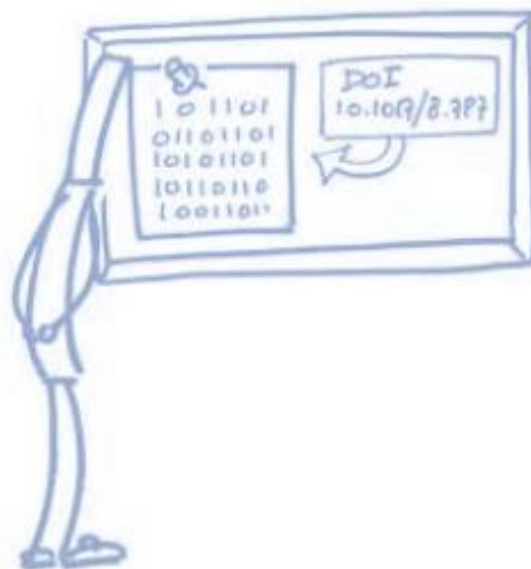
- Como lidar com o crescente volume de dados que surge a partir das mais variadas técnicas e estudos?
- Como acessar, organizar, gerenciar e processar estes grandes conjuntos de dados?
- Como explorar totalmente o potencial dos conjuntos de dados, especialmente em situações de grande demanda computacional?
- Como comparar novos dados com estudos prévios e permitir que esses dados sejam comparados com estudos futuros?

Princípios FAIR

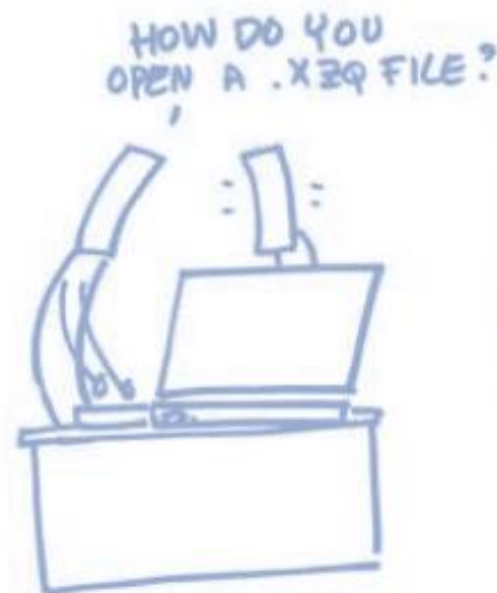
FAIR DATA PRINCIPLES



FINDABLE



ACCESSIBLE



INTEROPERABLE



REUSABLE

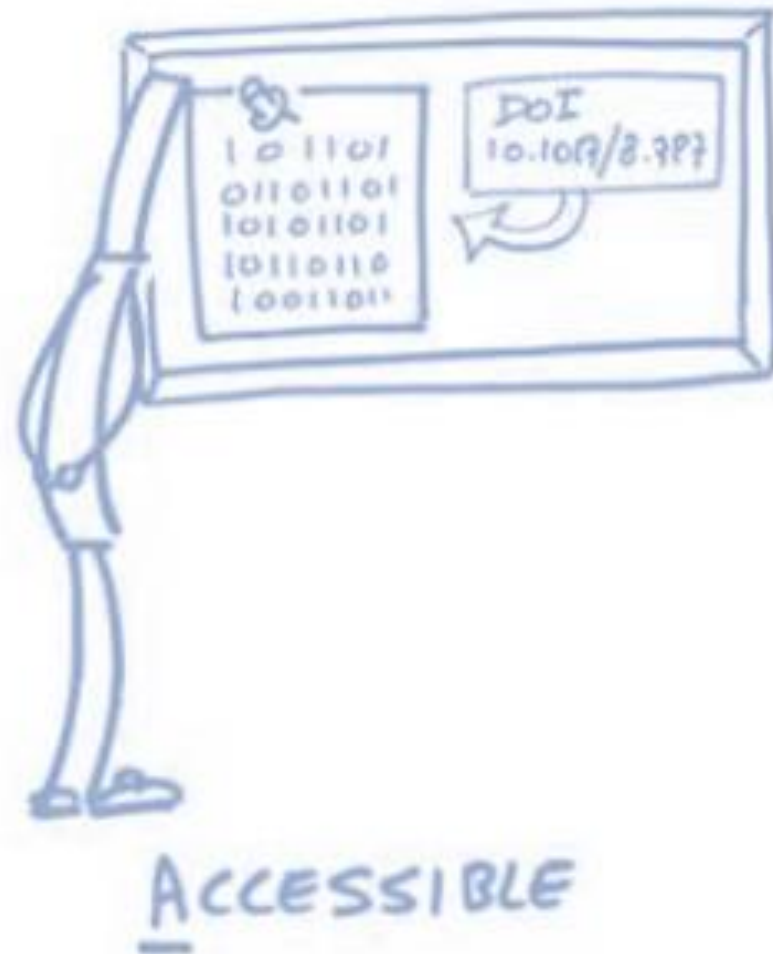
Findable / Localizáveis

- (Meta)dados devem ter identificadores globais, persistentes e identificáveis
- Dados devem ser descritos com metadados ricos
- (Meta)dados devem ser registrados e/ou indexados em recursos que permitam buscas
- Metadados devem especificar o identificador dos dados



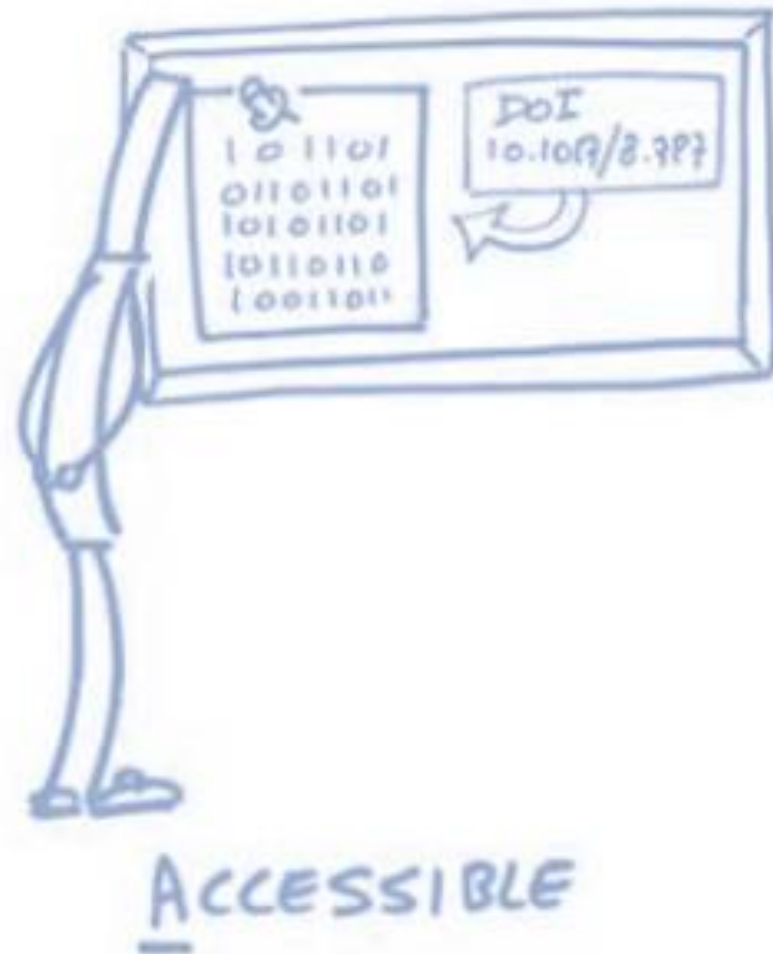
Acessible / Acessíveis

- (Meta)dados devem recuperáveis pelos seus identificadores usando um protocolo de comunicação padronizado
- O protocolo deve ser aberto, gratuito e universalmente implementável



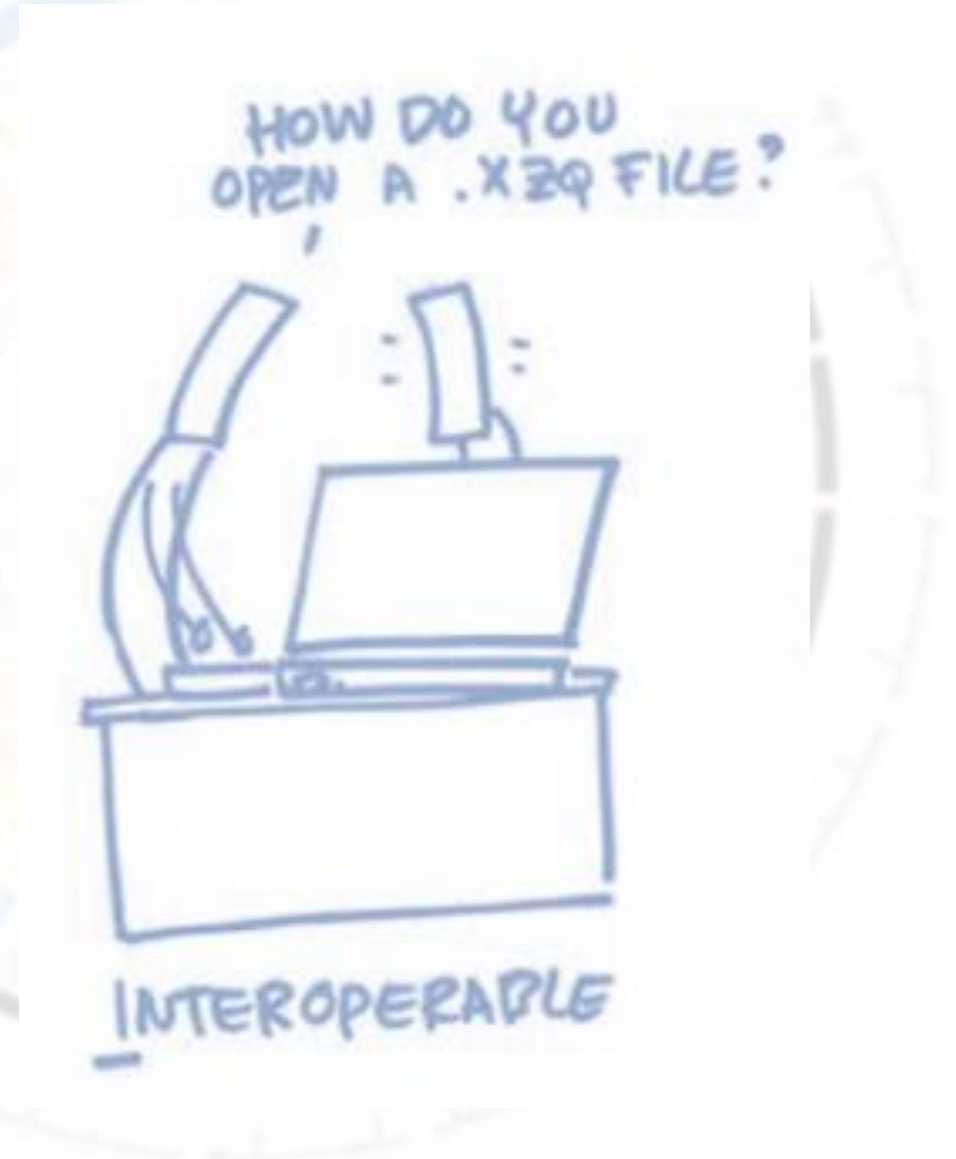
Acessible / Acessíveis

- O protocolo deve permitir procedimentos de autenticação e autorização quando necessário
- (Meta)dados devem ser acessíveis, mesmo quando os dados não estão mais disponíveis



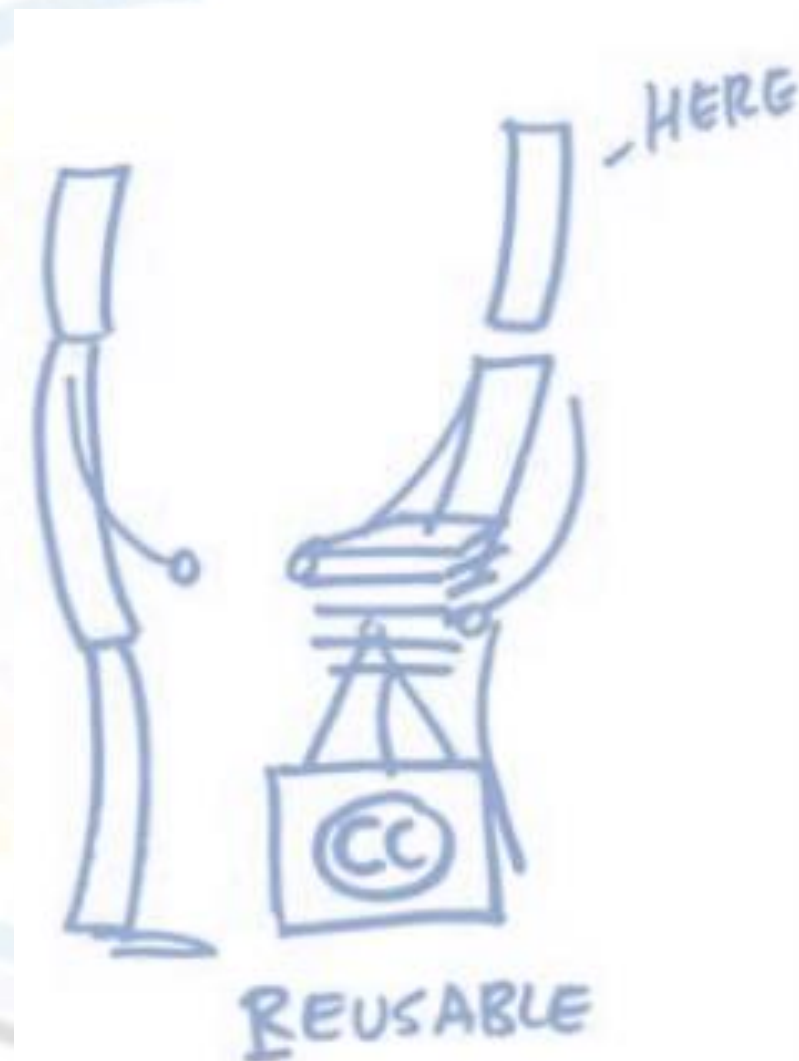
Interoperable / Interoperáveis

- (Meta)dados devem usar uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento
- (Meta)dados devem usar vocabulários que seguem os princípios FAIR
- (Meta)dados devem incluir referências qualificadas para outros (meta)dados



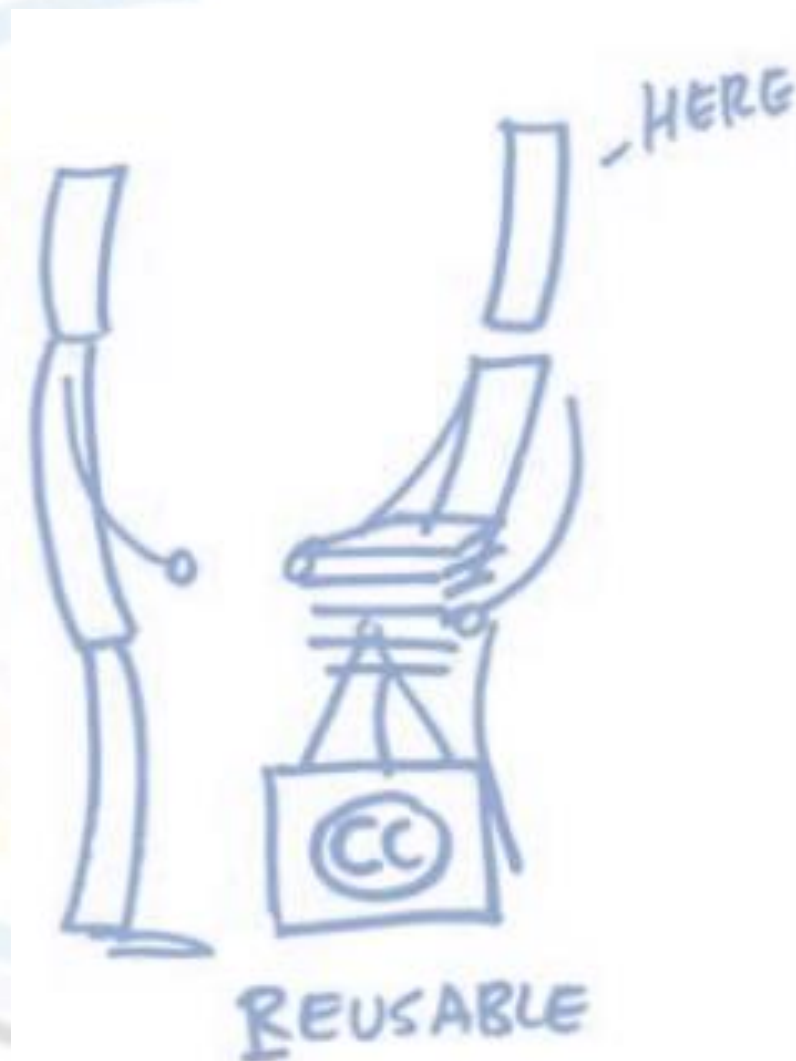
Reusable / Reutilizáveis

- (Meta)dados devem ter atributos com pluralidade de precisão e serem relevantes
- (Meta)dados devem ser liberados com licenças de uso de dados claras e acessíveis



Reusable / Reutilizáveis

- (Meta)dados devem estar associados à sua proveniência
- (Meta)dados devem estar alinhados com padrões relevantes ao seu domínio



	D	E
Jan	January	
Feb	Febuary	
Mar	Maruary	
Apr	Apruary	
May	Mayuary	
Jun	Junuary	
Jul	Juluary	
Aug	Auguary	
Sep	Sepuary	
Oct	Octuary	
Nov	Novuary	
Dec	Decuary	



jxf@mastodon.social
@jxxf

Optimist: The glass is $\frac{1}{2}$ full.

Pessimist: The glass is $\frac{1}{2}$ empty.

Excel: The glass is January 2nd



Correspondence

[Open Access](#)

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg^{†1}, Joseph Riss^{†2}, David W Kane³, Kimberly J Bussey¹, Edward Uchio⁴, W Marston Linehan⁴, J Carl Barrett² and John N Weinstein^{*1}

excel.gene2date.xls

	gene names	internal date format	default date format		gene names	internal date format	default date format		gene names	internal date format	default date format
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep

COMMENT

Open Access



CrossMark

Gene name errors are widespread in the scientific literature

Mark Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2* (Septin 2) and *MARCH1* [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively. Furthermore, RIKEN identifiers were described to be automatically converted to floating point numbers (i.e. from accession '2310009E13' to '2.31E+13'). Since that report, we have uncovered further instances where gene symbols were converted to dates in supplementary data of recently published papers (e.g. '*SEPT2*' converted to '2006/09/02'). This suggests that gene name errors continue to be a problem in supplementary files accompanying articles. Inadvertent gene symbol conversion is problematic because these supplementary files are an important resource in the genomics community that are frequently reused. Our aim here is to raise awareness of the problem.

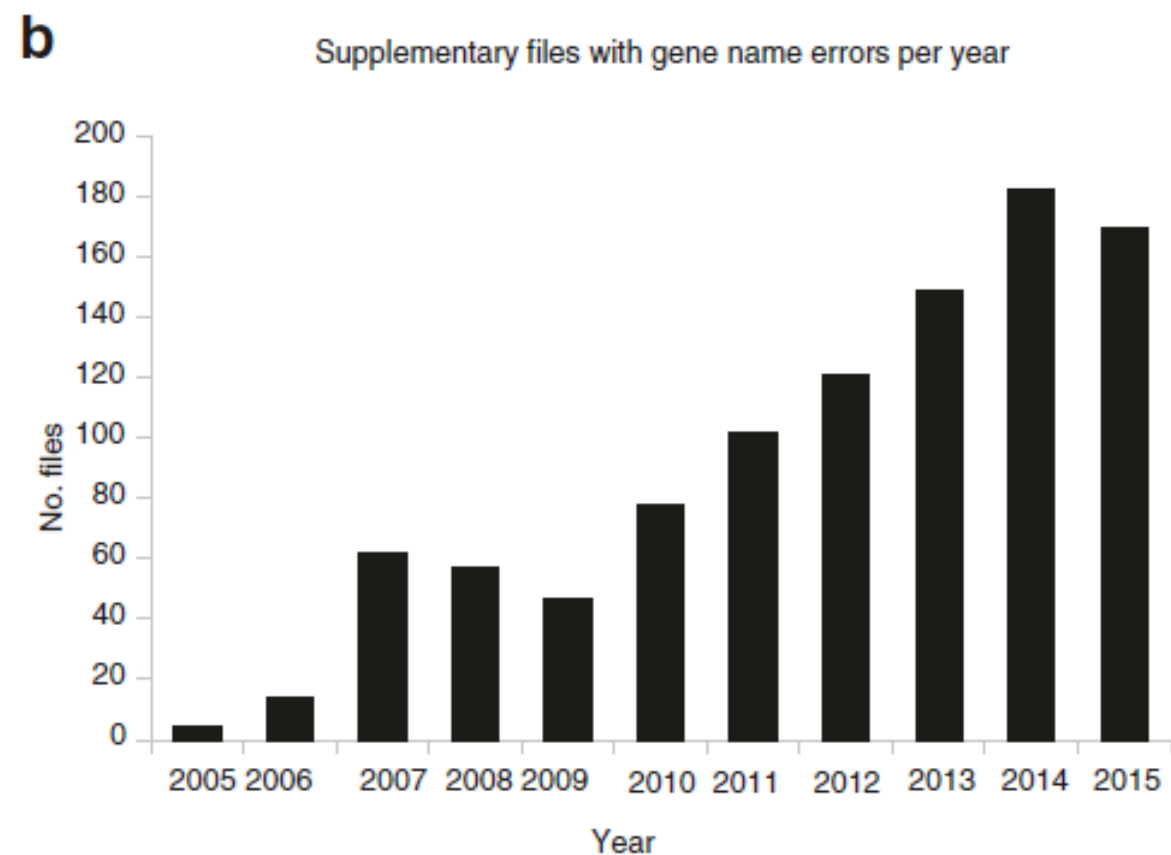
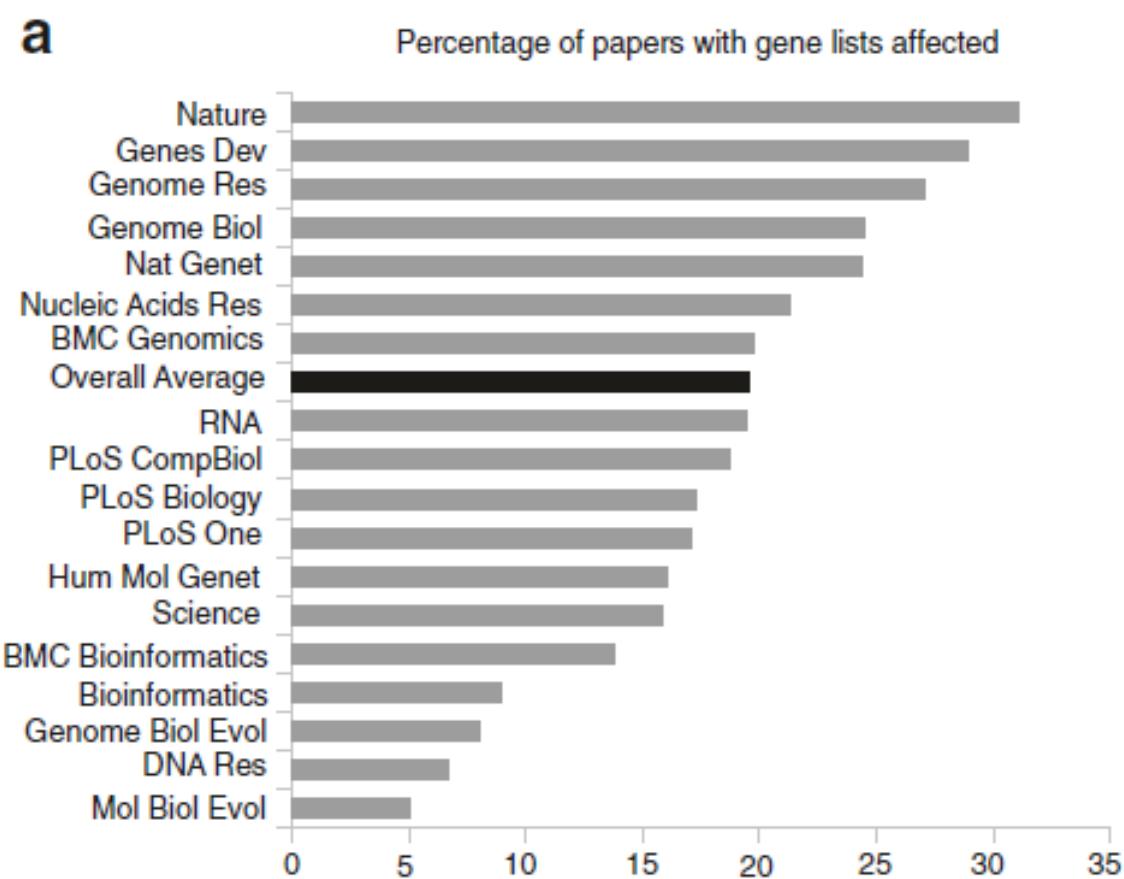


Fig. 1 Prevalence of gene name errors in supplementary Excel files. **a** Percentage of published papers with supplementary gene lists in Excel files affected by gene name errors. **b** Increase in gene name errors by year



Ireland Simpsons Fans

@iresimpsonsfans



The Microsoft Word doc after you move one image.



4:28 PM · Jun 26, 2021



Flat text file (texto simples)

- Informação armazenada em arquivo de texto simples
- Texto é estruturado (em função do tipo de dado armazenado)
- Vantagens:
 - Maior legibilidade
 - Possibilidade de acesso automático (especialmente importante no contexto de scripts e análises)
 - Fácil leitura por humanos e máquinas
- Exemplos: CSV, TXT, FASTA, GenBank, GFF3

Flat text file (texto simples)

```
C:\Users\Acer\Documents\Google Drive\2011-2018 - BioGeMM\2018-2022 - Phd\Gene Annotation\GFF3 files from EVM\PCapitalensis_CBS173_77.gff3 - Notepad++

Arquivo  Editar  Localizar  Visualizar  Formatar  Linguagem  Configurações  Ferramentas  Macro  Executar  Plugins  Janela  ?

PCapitalensis_CBS173_77.gff3

1 scaffold_10 EVM gene 2697 4438 . + . ID=evm.TU.scaffold_10.1;Name=EVM%20prediction%20scaffold_10.1
2 scaffold_10 EVM mRNA 2697 4438 . + . ID=evm.model.scaffold_10.1;Parent=evm.TU.scaffold_10.1;Name=EVM%20prediction%20scaffold_10.1
3 scaffold_10 EVM exon 2697 3078 . + . ID=evm.model.scaffold_10.1.exon1;Parent=evm.model.scaffold_10.1
4 scaffold_10 EVM CDS 2697 3078 . + 0 ID=cds.evm.model.scaffold_10.1;Parent=evm.model.scaffold_10.1
5 scaffold_10 EVM exon 3145 4023 . + . ID=evm.model.scaffold_10.1.exon2;Parent=evm.model.scaffold_10.1
6 scaffold_10 EVM CDS 3145 4023 . + 2 ID=cds.evm.model.scaffold_10.1;Parent=evm.model.scaffold_10.1
7 scaffold_10 EVM exon 4068 4438 . + . ID=evm.model.scaffold_10.1.exon3;Parent=evm.model.scaffold_10.1
8 scaffold_10 EVM CDS 4068 4438 . + 2 ID=cds.evm.model.scaffold_10.1;Parent=evm.model.scaffold_10.1
9
```

```
PCapitalensis_CBS173_77.gff3 - Bloco de Notas

Arquivo  Editar  Formatar  Exibir  Ajuda

scaffold_10 EVM gene 2697 4438 . + . ID=evm.TU.scaffold_10.1;Name=EVM%20prediction%20scaffold_10.1
scaffold_10 EVM mRNA 2697 4438 . + . ID=evm.model.scaffold_10.1;Parent=evm.TU.scaffold_10.1;Name=EVM%20prediction%20scaffold_10.1
scaffold_10 EVM exon 2697 3078 . + . ID=evm.model.scaffold_10.1.exon1;Parent=evm.model.scaffold_10.1
scaffold_10 EVM CDS 2697 3078 . + 0 ID=cds.evm.model.scaffold_10.1;Parent=evm.model.scaffold_10.1
scaffold_10 EVM exon 3145 4023 . + . ID=evm.model.scaffold_10.1.exon2;Parent=evm.model.scaffold_10.1
scaffold_10 EVM CDS 3145 4023 . + 2 ID=cds.evm.model.scaffold_10.1;Parent=evm.model.scaffold_10.1
scaffold_10 EVM exon 4068 4438 . + . ID=evm.model.scaffold_10.1.exon3;Parent=evm.model.scaffold_10.1
scaffold_10 EVM CDS 4068 4438 . + 2 ID=cds.evm.model.scaffold_10.1;Parent=evm.model.scaffold_10.1
```


[illegible]

Formato FASTA

- **Extensões:** .fasta, .fas, .fna, .ffn, .faa, .frn, .fa
- **Linha 1:** identificador da sequência após o sinal de maior (>)
- **Linha 2 até que o próximo sinal de maior seja encontrado:** sequência
- Pode conter sequências de DNA ou proteínas

```
1 >scaffold_1
2 CCATGGCTGTCTTGCGATTGTCCAGGGCAGTCTTGACAGCAGGGGCAAGTTGCGCCGCCGCCCGTCCCTT
3 CTCAGTGTCTTCGAAGTTGAGGGAGACGATGACCCTGGTGTGATGGGACTGTTGGTGTTCGCCGTGGAA
4 GCTTCGTCCTTCTTGCGCTTGGAGCCGGCCGACGCCGCCTTCTTGCGCTTGGCAATCTCCTCGGGATGCG
5 TGAGAATCTCTTCAATTTTTGCCATGAAGGCGTTCTCTTTCTCGATTTACGAGCGAACTTCCTCGTAGAA
6 TGCCGTGGCTACGCTTGACTCGGTCTTGAAGATGTTGTGGAGAGCCTTGCGGGTGGTCGTTTACGACGAA
7 GATGCAGAGAGGGCGCGGTCTGCGCGATGGCGTTGCGGGTGTCTTCGTCTTGTTGAACCAGT
8 TGCCGAACCTGAATTACGTCGTCTTTCTTTGCCATCTTTTCCTCGGAGCTCATCGCTTCGATGGTGGCGGC
9 GTCATCCTTGCGCTTTTCGGCGGTCTCGTTTTTTCTTCGTCTGGGTGACTTGCAGAAGTGCCTTTGCCCTC
10 AAAGCACTCATTCGGCGACTCTCCTGTTCCGCATCGACGACCTGGCGCCATTCCTGTGACGCATCGCTCA
```

Formato GenBank

- **Extensão: .gbk**
- **Header (Cabeçalho)**
 - Código de acesso
 - Organismo
 - Publicações associadas

1	LOCUS	NC_045512	29903 bp ss-RNA	linear	VRL 18-JUL-2020
2	DEFINITION	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.			
3					
4	ACCESSION	NC_045512			
5	VERSION	NC_045512.2			
6	DBLINK	BioProject: PRJNA485481			
7	KEYWORDS	RefSeq.			
8	SOURCE	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)			
9	ORGANISM	Severe acute respiratory syndrome coronavirus 2			
10		Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes;			
11		Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae;			
12		Betacoronavirus; Sarbecovirus.			
13	REFERENCE	1 (bases 1 to 29903)			
14	AUTHORS	Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y.,			
15		Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H.,			
16		Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C. and Zhang,Y.Z.			
17	TITLE	A new coronavirus associated with human respiratory disease in China			
18					
19	JOURNAL	Nature 579 (7798), 265-269 (2020)			
20	PUBMED	32015508			
21	REMARK	Erratum:[Nature. 2020 Apr;580(7803):E7. PMID: 32296181]			

Formato GenBank

- **Extensão: .gbk**
- **Feature Table (Tabela de características)**
 - Informações sobre o indivíduo sequenciado
 - Anotações na sequência: gene, mRNA, CDS, regiões 5' e 3' UTR, sequência de aminoácidos de proteínas presentes na sequência, domínios conservados

72	FEATURES	Location/Qualifiers
73	source	1..29903
74		/organism="Severe acute respiratory syndrome coronavirus 2"
75		
76		/mol_type="genomic RNA"
77		/isolate="Wuhan-Hu-1"
78		/host="Homo sapiens"
79		/db_xref="taxon:2697049"
80		/country="China"
81		/collection_date="Dec-2019"
82	5'UTR	1..265
83	gene	266..21555
84		/gene="ORF1ab"
85		/locus_tag="GU280_gp01"
86		/db_xref="GeneID:43740578"
87	CDS	join(266..13468,13468..21555)
88		/gene="ORF1ab"
89		/locus_tag="GU280_gp01"
90		/ribosomal_slippage
91		/note="pp1ab; translated by -1 ribosomal frameshift"
92		/codon_start=1
93		/product="ORF1ab polyprotein"
94		/protein_id="YP_009724389.1"
95		/db_xref="GeneID:43740578"
96		/translation="MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQ
97		HLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGE
98		TLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDDELGTDPYEDFQEN
99		WNTKHSSGVTRELMRELNGGAYTRYVDNNEFCGPDGYPLECIKDLLARAGKASCTLSEQ
100		LDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFP
101		LNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTG

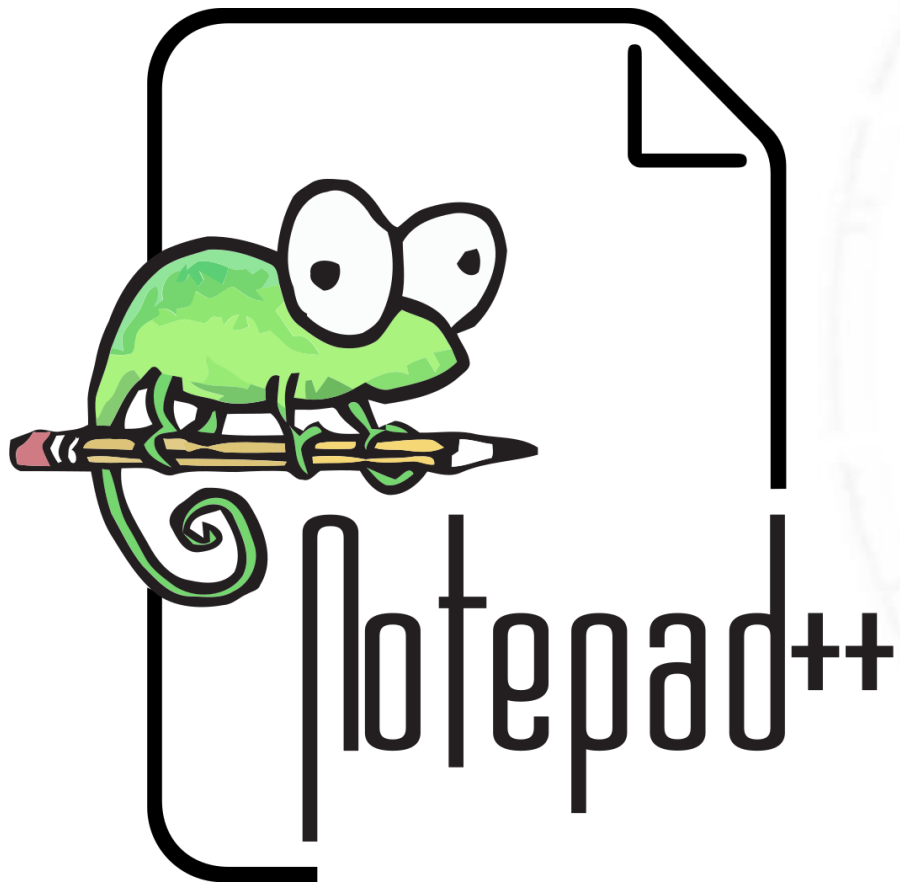
Formato GenBank

- **Extensão: .gbk**
- **Origin**
 - Sequência de DNA ou proteína propriamente dita
- **Final do arquivo**
 - Sempre termina com "//"

```
690 ORIGIN
691      1 attaaagggt tataccttcc caggtaacaa accaaccaac tttcgatctc ttgtagatct
692     61 gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgactc
693    121 cacgcagtat aattaataac taattactgt cgttgacagg acacgagtaa ctctgtctatc
694    181 ttctgcaggc tgcttacggg ttcgtccgtg ttgcagccga tcatcagcac atctagggtt
695    241 cgtccgggtg tgaccgaaag gtaagatgga gagccttgtc cctggtttca acgagaaaac
696    301 acacgtccaa ctacgtttgc ctgttttaca ggttcgcgac gtgctcgtac gtggccttgg
697    361 agactccgtg gaggagggtc tatcagaggc acgtcaacat cttaaagatg gcacttgtgg
698    421 cttagtagaa gttgaaaaag gcgttttgcc tcaacttgaa cagccctatg tgttcatcaa
699    481 acgttcggat gtcgaactg cacctcatgg tcatgttatg gttgagctgg tagcagaact
700    541 cgaaggcatt cagtacgggc gtagtggtga gacacttggt gtccttgtcc ctcatgtggg
```

```
1179    29281 catcaaattg gatgacaaag atccaaattt caaagatcaa gtcattttgc tgaataagca
1180    29341 tattgacgca tacaaaacat tcccaccaac agagcctaaa aaggacaaaa agaagaaggc
1181    29401 tgatgaaact caagccttac cgcagagaca gaagaaacag caaactgtga ctcttcttcc
1182    29461 tgctgcagat ttggatgatt tctccaaaca attgcaacaa tccatgagca gtgctgactc
1183    29521 aactcaggcc taaactcatg cagaccacac aaggcagatg ggctatataa acgttttcgc
1184    29581 ttttccggtt acgatataa gtctactctt gtgcagaatg aattctcgta actacatagc
1185    29641 acaagtagat gtagttaact ttaatctcac atagcaatct ttaatcagtg tgtaacatta
1186    29701 gggaggactt gaaagagcca ccacattttc accgaggcca cgcgaggtac gatcgagtgt
1187    29761 acagtgaaca atgctaggga gagctgccta tatggaagag ccctaattgtg taaaattaat
1188    29821 tttagtagtg ctatcccat gtgattttta tagcttctta ggagaatgac aaaaaaaaaa
1189    29881 aaaaaaaaaa aaaaaaaaaa aaa
1190 //
1191
```

Sugestões – Editores de Texto



Por que trabalhar com softwares livres?

- Possibilidades para o usuário:
 - Executar o software para qualquer propósito
 - Estudar o software e adaptar conforme necessidades
 - Redistribuir (mantendo a configuração original)
 - Modificar e redistribuir