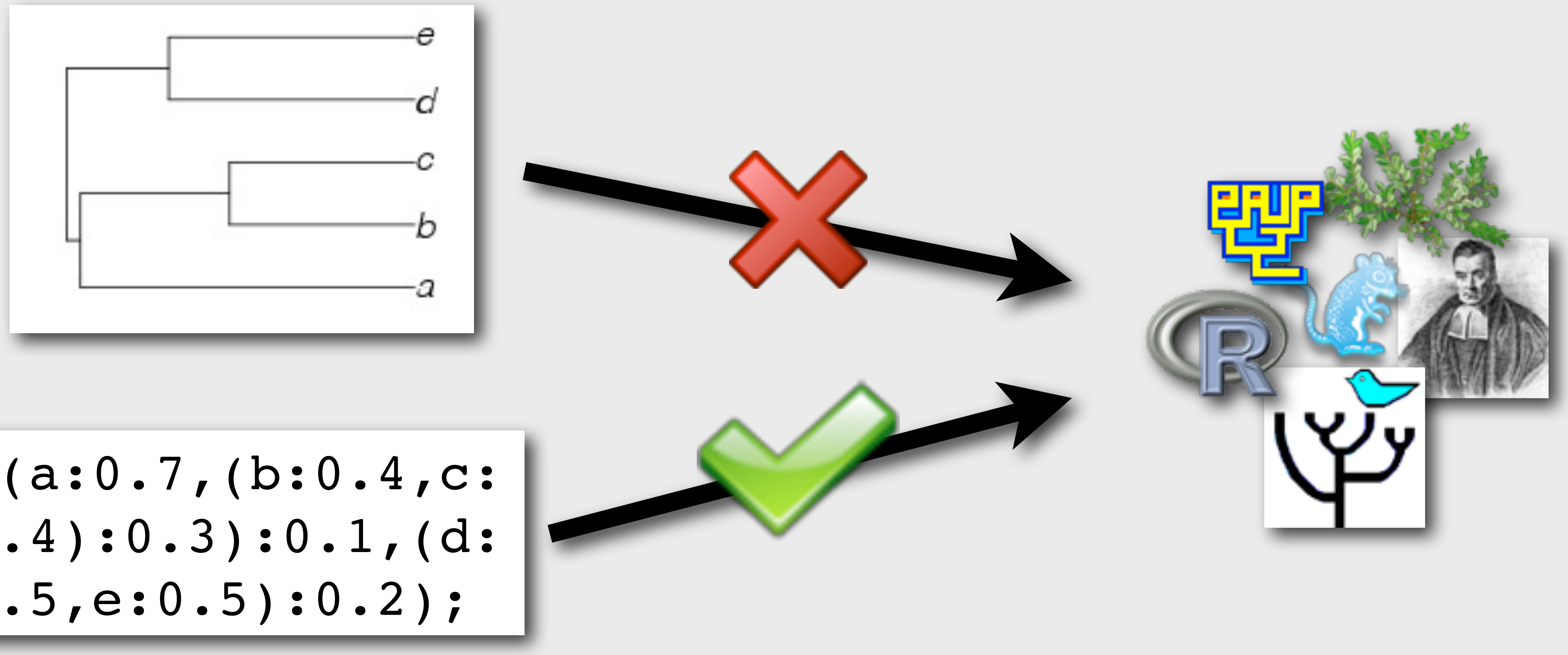
Current and best practices for sharing trees



```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

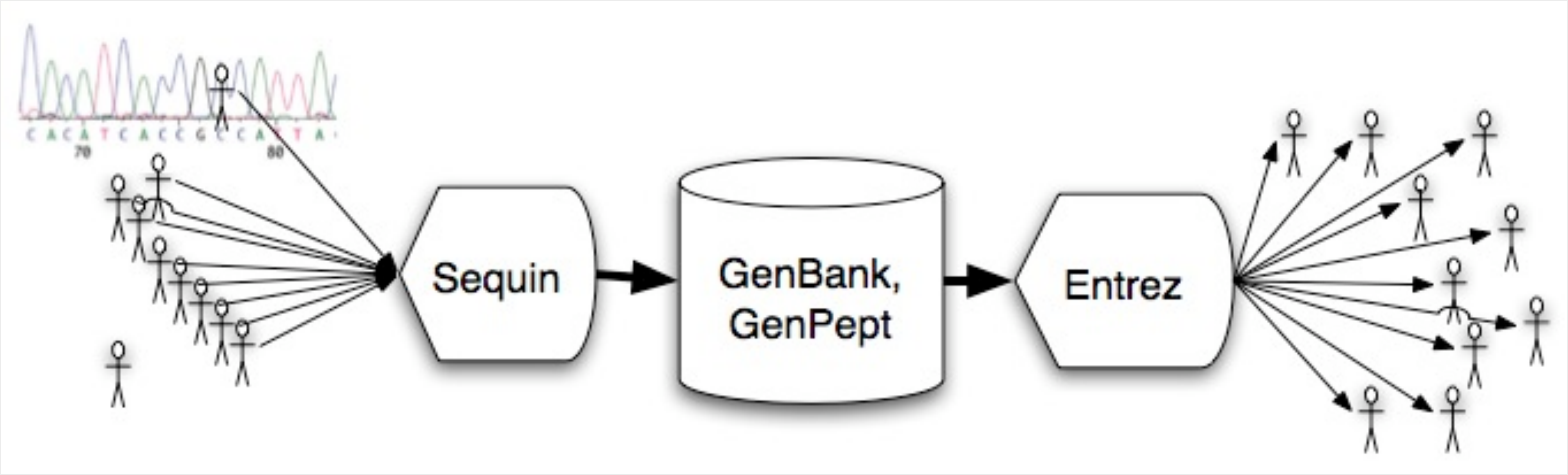
Arlin Stoltzfus, Brian O'Meara, Jamie Whitacre, Ross Mounce, Dan Rosauer

Tuesday, November 1, 2011

Modern information technology creates an enormous potential for sharing and re-use of scientific data, including comparative data and phylogenies. We have the technology to ensure that scientists today can re-use and build on all the phylogenies that other scientists made yesterday, or this morning.

In practice, the success of sharing depends no only on technology, but also on standards and on community practices— the focus of my talk today.

Sharing sequence data



Producers

Submission tool

Repository

Search & retrieval tool

Consumers (re-users)

Tuesday, November 1, 2011

To understand this complex challenge, it helps to have a point of reference, and my point of reference is going to be sharing molecular sequence data. How do scientists share sequence data? Why isn't it as easy to share phylogenetic data as it is to share sequence data?

On the left, we have the producers of new data, who often are the primary consumers in the sense that they not only generate data, but interpret it and apply it to scientific questions. On the right we have secondary consumers or re-users of data. For DNA sequence data, sharing is mediated by a public archive, GenBank. Using a convenient submission tool called "Sequin", producers archive their sequence data in GenBank. Consumers can access the data via a convenient and powerful web-based search-and-retrieval interface called "Entrez".

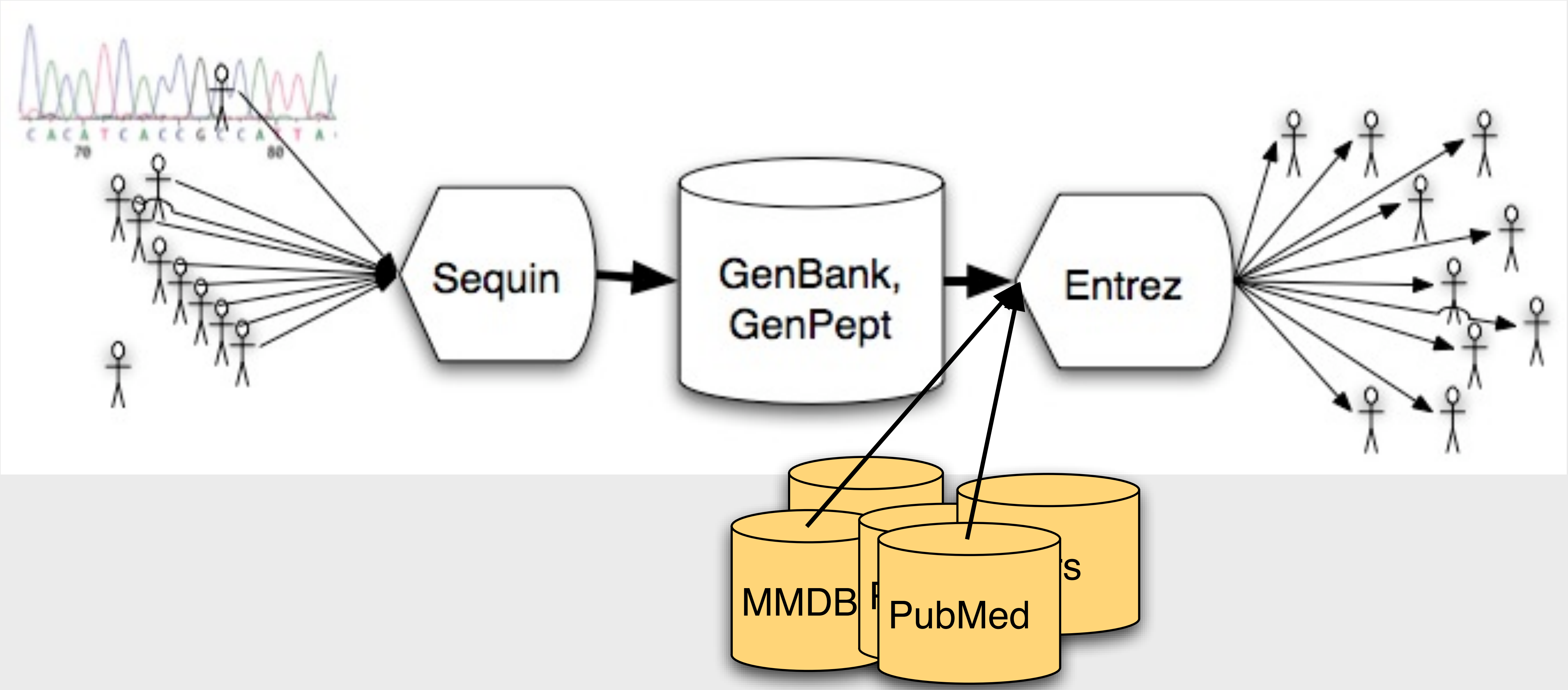
click: To complete this picture, we may note that the value Entrez as a discovery tool is greatly enhanced by its access to links to other kinds of data, such as publications in PubMed.

click: Consumers also have other interfaces such as BLAST. Power users have access to "eutils" (web services) and FTP downloads of sequence data.

click: Secondary value-added resources such as Pandit, Pfam and COGS are computed from the data in GenBank. They represent additional conduits for archive-mediated sharing of sequence data.

[notes. There was a time when protein sequences were determined directly and deposited in GenPept, but now nearly all protein sequences are inferred from DNA sequences.]

Sharing sequence data



Producers

Repository

Consumers (re-users)

Submission tool

Search & retrieval tool

To understand this complex challenge, it helps to have a point of reference, and my point of reference is going to be sharing molecular sequence data. How do scientists share sequence data? Why isn't it as easy to share phylogenetic data as it is to share sequence data?

On the left, we have the producers of new data, who often are the primary consumers in the sense that they not only generate data, but interpret it and apply it to scientific questions. On the right we have secondary consumers or re-users of data. For DNA sequence data, sharing is mediated by a public archive, GenBank. Using a convenient submission tool called "Sequin", producers archive their sequence data in GenBank. Consumers can access the data via a convenient and powerful web-based search-and-retrieval interface called "Entrez".

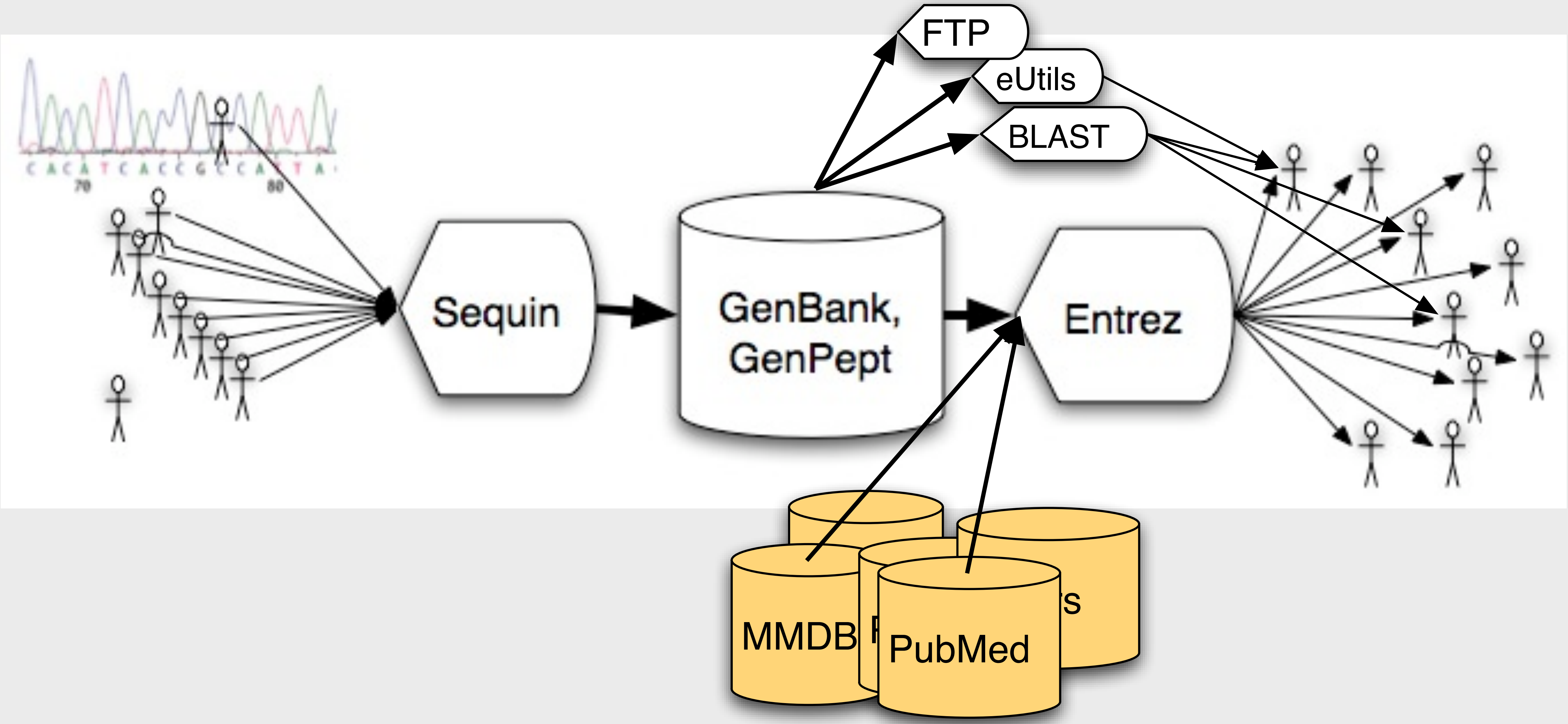
click: To complete this picture, we may note that the value Entrez as a discovery tool is greatly enhanced by its access to links to other kinds of data, such as publications in PubMed.

click: Consumers also have other interfaces such as BLAST. Power users have access to "eutils" (web services) and FTP downloads of sequence data.

click: Secondary value-added resources such as Pandit, Pfam and COGS are computed from the data in GenBank. They represent additional conduits for archive-mediated sharing of sequence data.

[notes. There was a time when protein sequences were determined directly and deposited in GenPept, but now nearly all protein sequences are inferred from DNA sequences.]

Sharing sequence data



Producers

Repository

Consumers (re-users)

Submission tool

Search & retrieval tool

To understand this complex challenge, it helps to have a point of reference, and my point of reference is going to be sharing molecular sequence data. How do scientists share sequence data? Why isn't it as easy to share phylogenetic data as it is to share sequence data?

On the left, we have the producers of new data, who often are the primary consumers in the sense that they not only generate data, but interpret it and apply it to scientific questions. On the right we have secondary consumers or re-users of data. For DNA sequence data, sharing is mediated by a public archive, GenBank. Using a convenient submission tool called "Sequin", producers archive their sequence data in GenBank. Consumers can access the data via a convenient and powerful web-based search-and-retrieval interface called "Entrez".

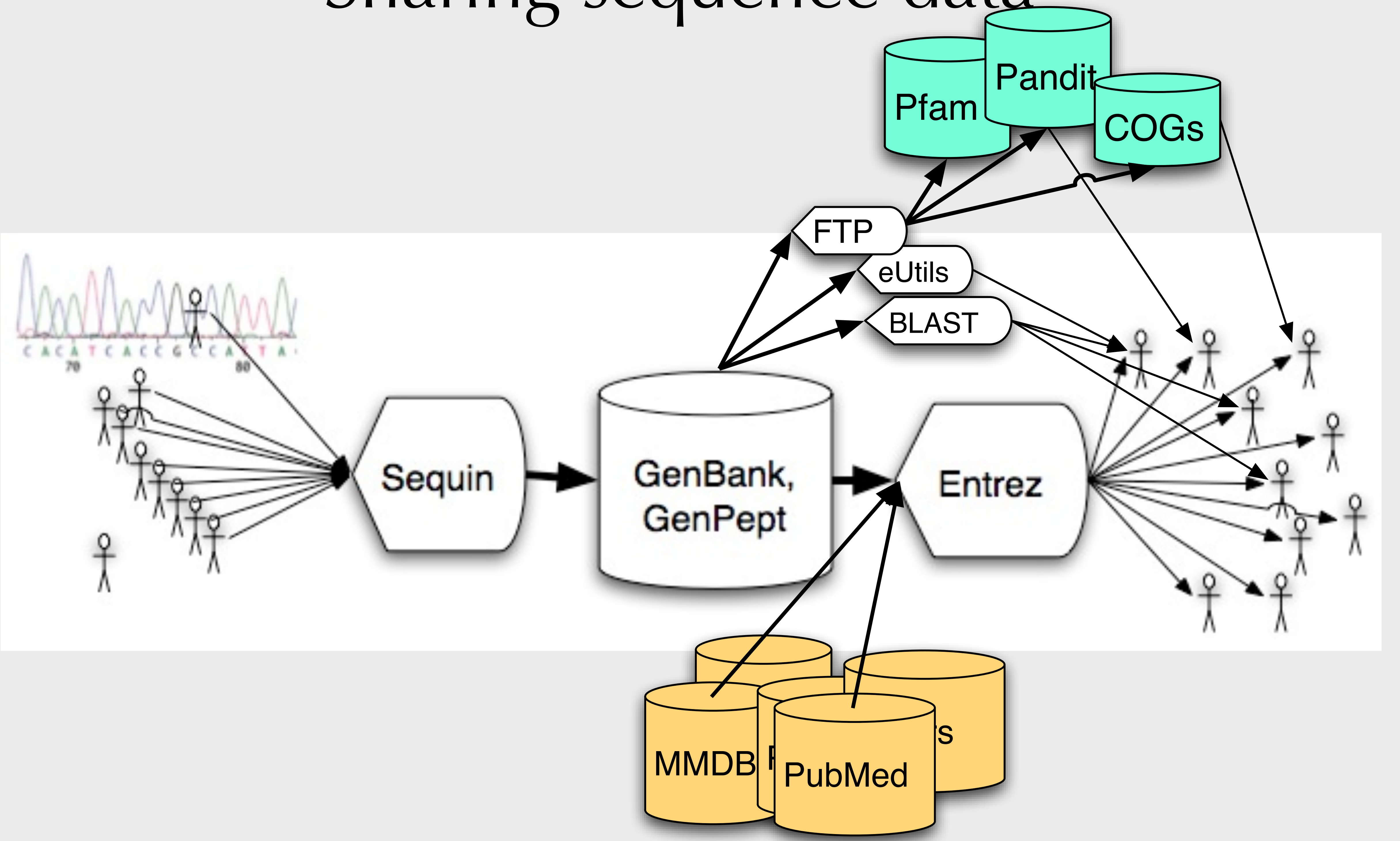
click: To complete this picture, we may note that the value Entrez as a discovery tool is greatly enhanced by its access to links to other kinds of data, such as publications in PubMed.

click: Consumers also have other interfaces such as BLAST. Power users have access to "eutils" (web services) and FTP downloads of sequence data.

click: Secondary value-added resources such as Pandit, Pfam and COGS are computed from the data in GenBank. They represent additional conduits for archive-mediated sharing of sequence data.

[notes. There was a time when protein sequences were determined directly and deposited in GenPept, but now nearly all protein sequences are inferred from DNA sequences.]

Sharing sequence data



Pfam

Pandit

COGs

FTP

eUtils

BLAST

Sequin

GenBank, GenPept

Entrez

MMDB

PubMed

Producers

Repository

Consumers (re-users)

Submission tool

Search & retrieval tool

To understand this complex challenge, it helps to have a point of reference, and my point of reference is going to be sharing molecular sequence data. How do scientists share sequence data? Why isn't it as easy to share phylogenetic data as it is to share sequence data?

On the left, we have the producers of new data, who often are the primary consumers in the sense that they not only generate data, but interpret it and apply it to scientific questions. On the right we have secondary consumers or re-users of data. For DNA sequence data, sharing is mediated by a public archive, GenBank. Using a convenient submission tool called "Sequin", producers archive their sequence data in GenBank. Consumers can access the data via a convenient and powerful web-based search-and-retrieval interface called "Entrez".

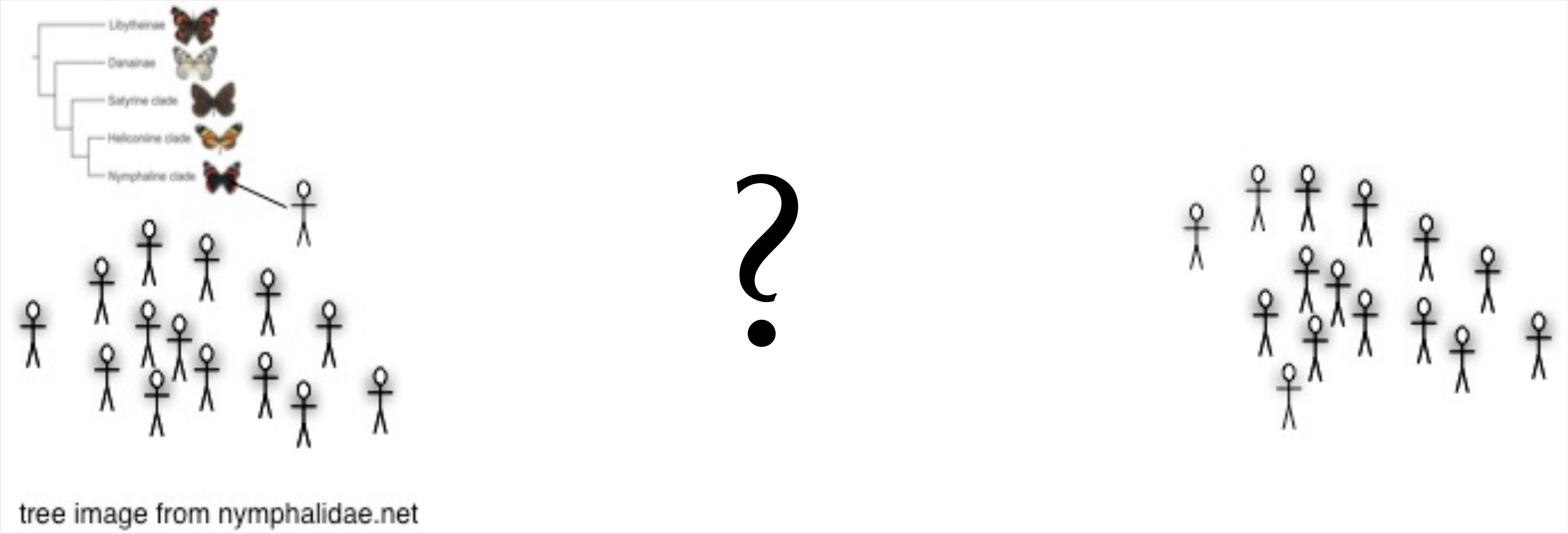
click: To complete this picture, we may note that the value Entrez as a discovery tool is greatly enhanced by its access to links to other kinds of data, such as publications in PubMed.

click: Consumers also have other interfaces such as BLAST. Power users have access to "eutils" (web services) and FTP downloads of sequence data.

click: Secondary value-added resources such as Pandit, Pfam and COGS are computed from the data in GenBank. They represent additional conduits for archive-mediated sharing of sequence data.

[notes. There was a time when protein sequences were determined directly and deposited in GenPept, but now nearly all protein sequences are inferred from DNA sequences.]

Sharing trees



tree image from nymphalidae.net

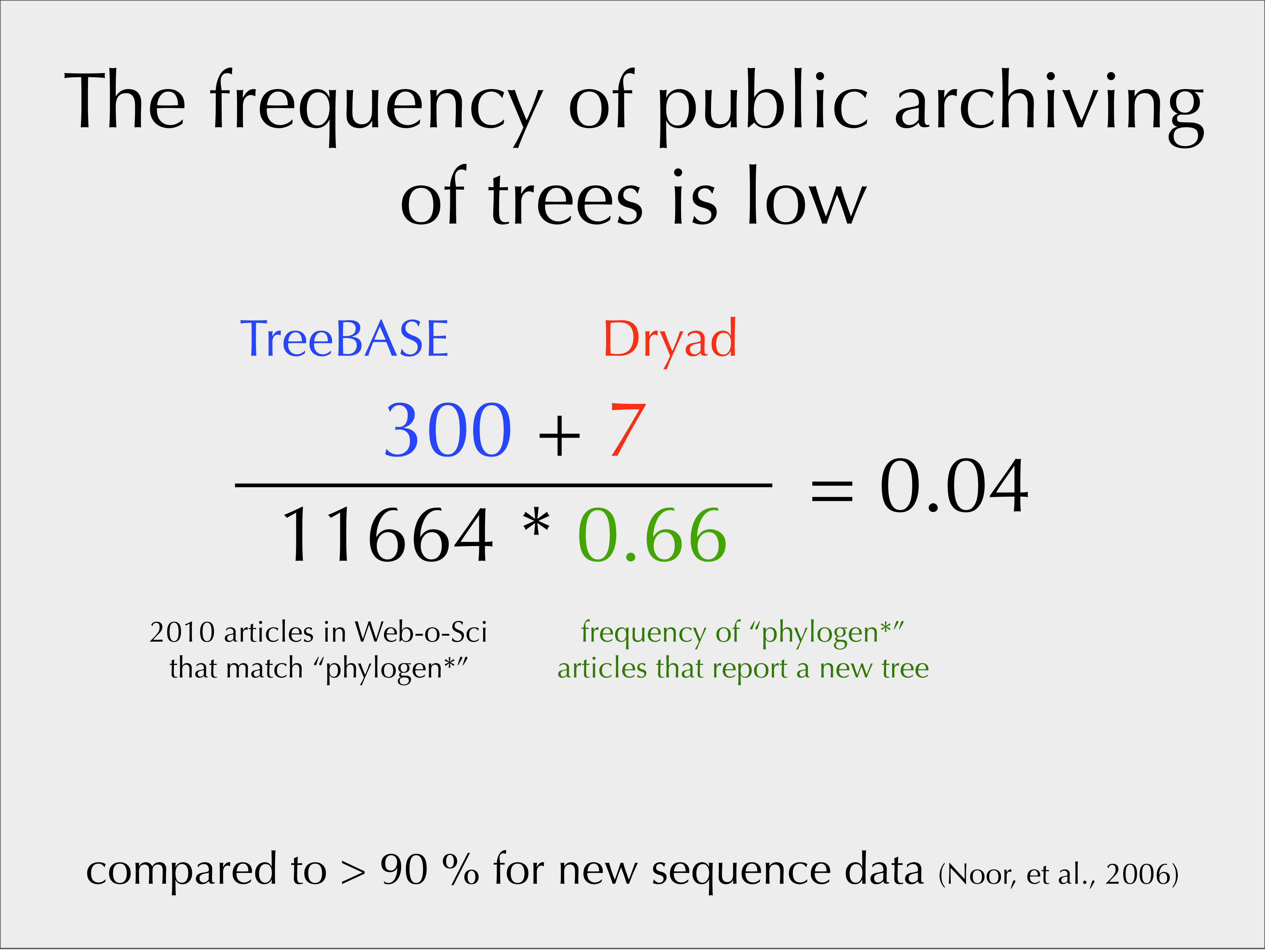
Producers

Consumers (re-users)

Tuesday, November 1, 2011

(brief, just a few seconds) How would we depict the overall economy of sharing for phylogenetic trees and associated data? What kinds of data are being shared? What practices and technologies are used in sharing?   How often does sharing occur?

After building up a picture of current practices, and then consider that picture from the perspective of how to facilitate sharing by lowering barriers.



The frequency of public archiving of trees is low

TreeBASE Dryad

$$\frac{300 + 7}{11664 * 0.66} = 0.04$$

2010 articles in Web-o-Sci that match “phylogen*”

frequency of “phylogen*” articles that report a new tree

compared to > 90 % for new sequence data (Noor, et al., 2006)

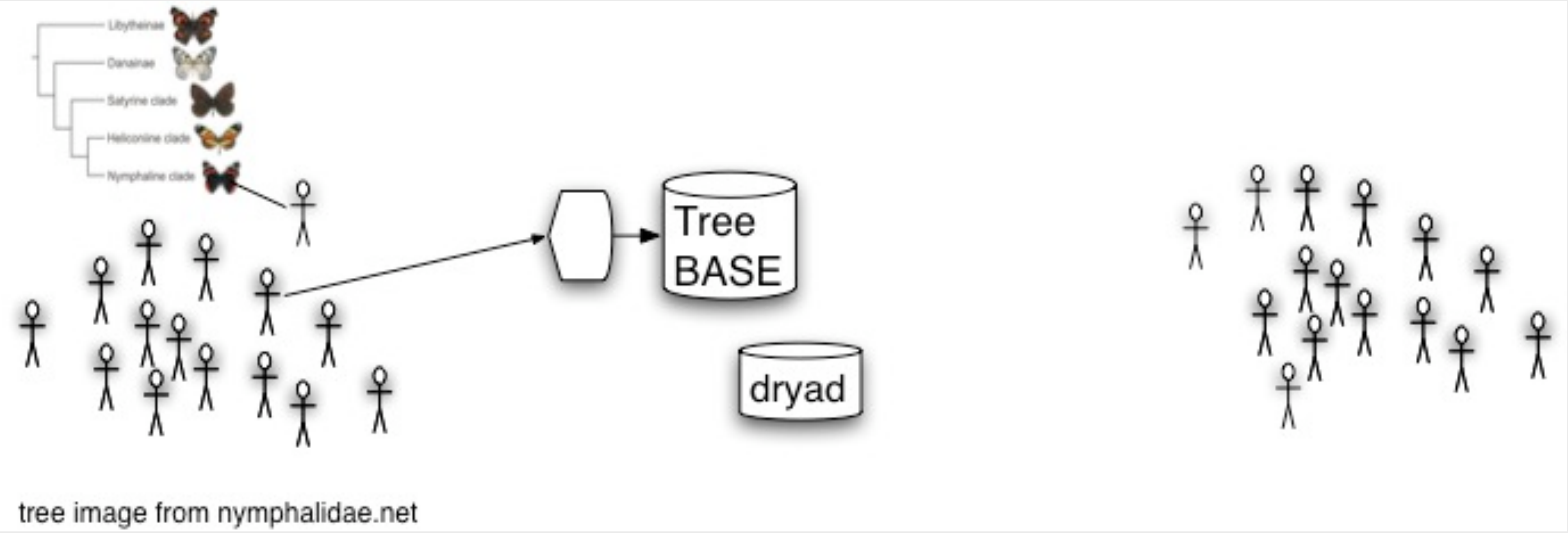
Tuesday, November 1, 2011

The first thing to note is the frequency of archiving trees. Let’s start with the denominator in this equation. We estimated the number of new phylogeny reports by searching for 2010 publications matching the string “phylogen*”, then correcting for false positives. The factor of 0.66 (green) is based on reading a random sample of 100 articles: only 66 actually report a new tree. In other words, 2/3 of papers that match “phylogen*” actually report a new tree. This suggests that there were about 7700 reports of phylogenies in 2010.

Meanwhile, there are a total of 307 archival records in TreeBASE and Dryad, for articles published in 2010. This gives an overall frequency of archiving of 4 %, in comparison to over 90 % for DNA sequences.

[notes: There may be other small archives not listed here. MorphoBank accepts matrices, but not trees. In 7700 articles, there may be many more than 7700 trees reported]

Sharing trees



tree image from nymphalidae.net

Producers

Repositories

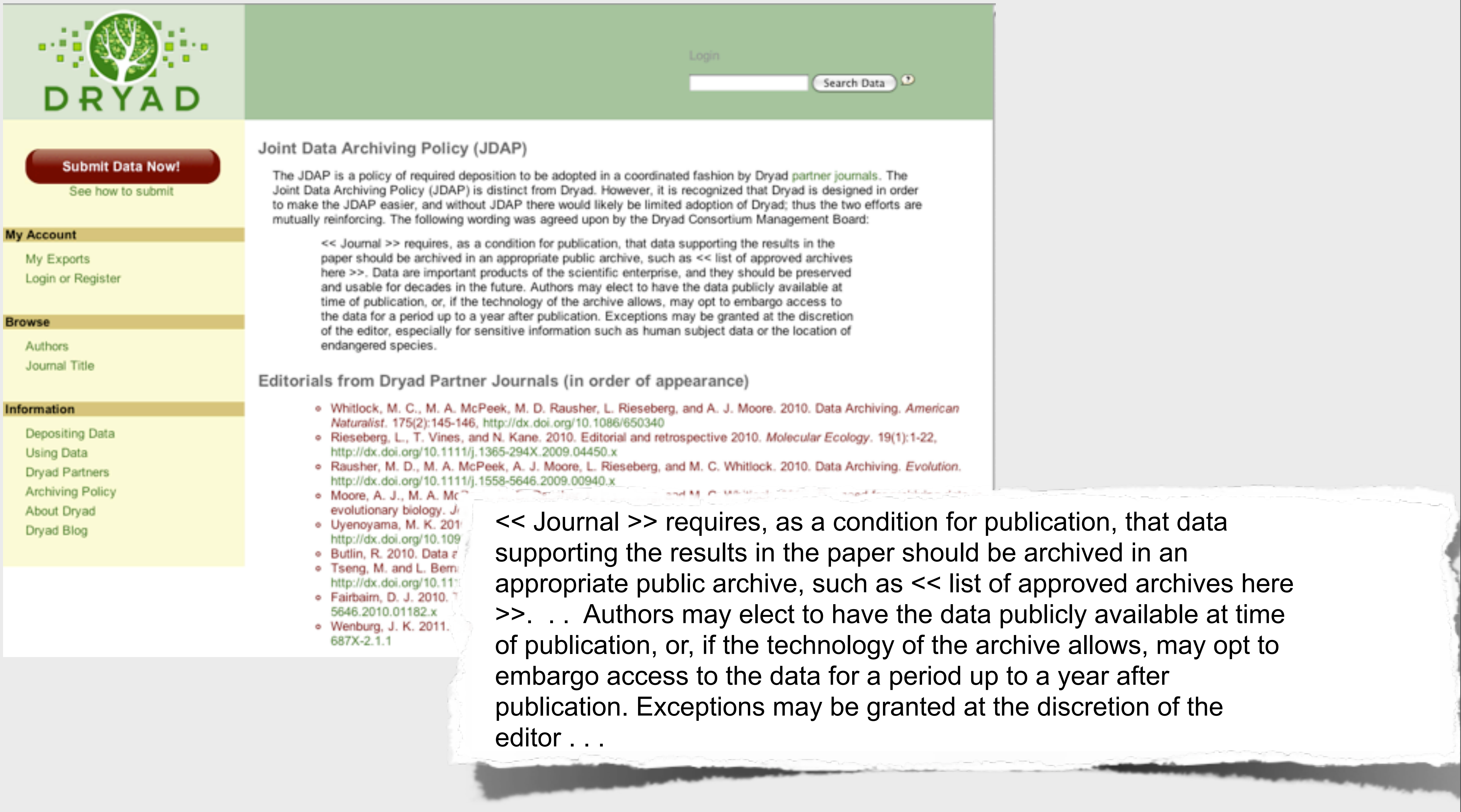
Consumers (re-users)

Tuesday, November 1, 2011

(brief, just a few seconds) So, the first thing we know is that the frequency of archiving in public archives is low.

Note that, just as GenBank provides a convenient submission tool that collects metadata for new sequences, there is an interactive submission tool for TreeBASE that facilitates submission of a data matrix, a phylogeny, and basic metadata.

What about the joint data archiving policy that took effect in 2011?



Joint Data Archiving Policy (JDAP)

The JDAP is a policy of required deposition to be adopted in a coordinated fashion by Dryad partner journals. The Joint Data Archiving Policy (JDAP) is distinct from Dryad. However, it is recognized that Dryad is designed in order to make the JDAP easier, and without JDAP there would likely be limited adoption of Dryad; thus the two efforts are mutually reinforcing. The following wording was agreed upon by the Dryad Consortium Management Board:

> << Journal >> requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as << list of approved archives here >>. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species.

Editorials from Dryad Partner Journals (in order of appearance)

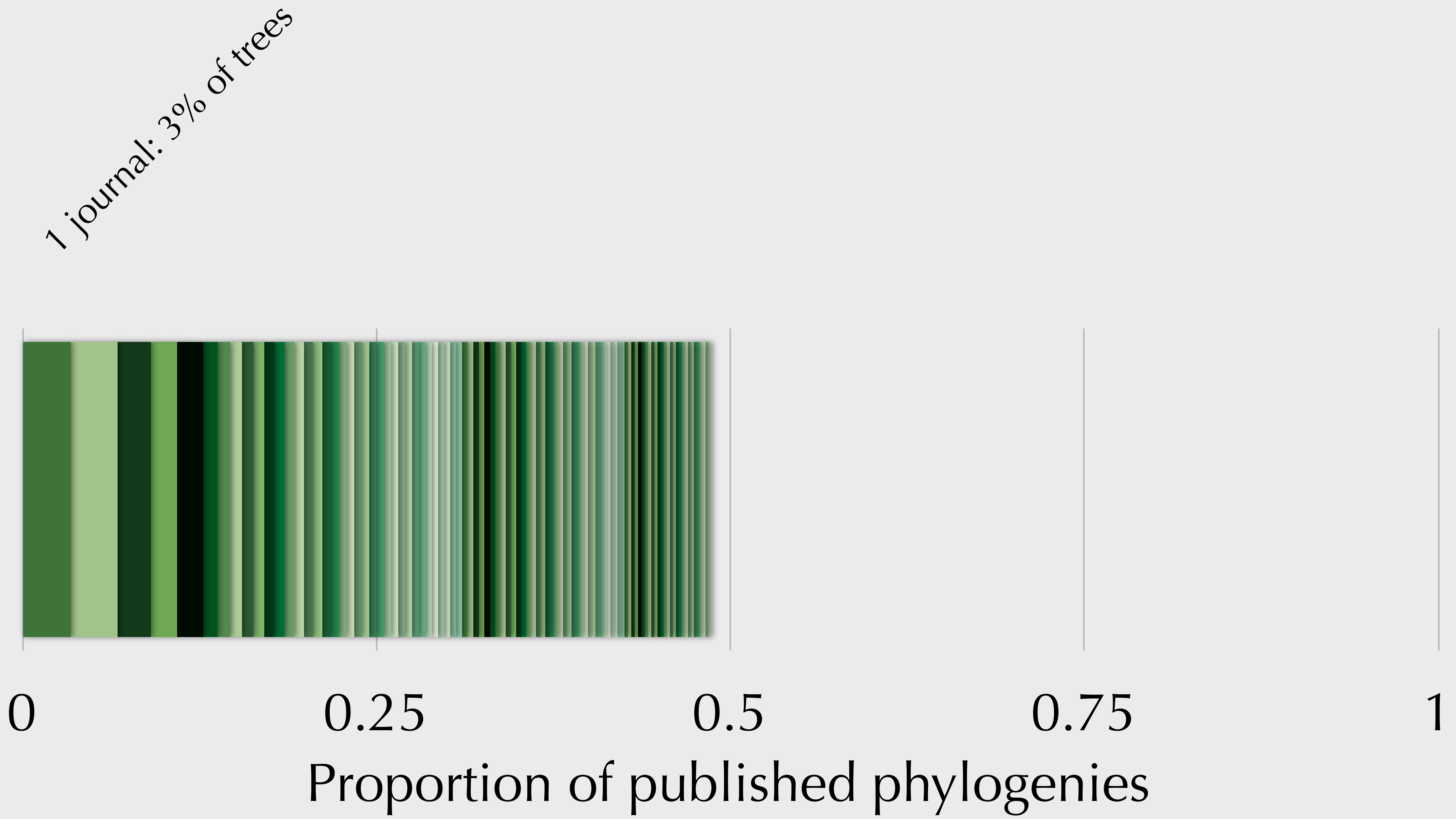
- Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data Archiving. *American Naturalist*. 175(2):145-146, http://dx.doi.org/10.1086/650340
- Rieseberg, L., T. Vines, and N. Kane. 2010. Editorial and retrospective 2010. *Molecular Ecology*. 19(1):1-22, http://dx.doi.org/10.1111/j.1365-294X.2009.04450.x
- Rausher, M. D., M. A. McPeek, A. J. Moore, L. Rieseberg, and M. C. Whitlock. 2010. Data Archiving. *Evolution*. http://dx.doi.org/10.1111/j.1558-5646.2009.00940.x
- Moore, A. J., M. A. Mc evolutionary biology. *J*
- Uyenoyama, M. K. 201 http://dx.doi.org/10.109
- Butlin, R. 2010. Data a
- Tseng, M. and L. Bern http://dx.doi.org/10.11
- Fairbairn, D. J. 2010. 5646.2010.01182.x
- Wenburg, J. K. 2011. 687X-2.1.1

<< Journal >> requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as << list of approved archives here >>. . . Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor . . .

Tuesday, November 1, 2011

(brief) That analysis was for articles published in 2010. Is it possible that archiving will take a dramatic jump, due to the Joint Data Archiving Policy that went into effect in January of 2011? We don't think so, for reasons explained on the next slide.

How are phylogeny articles distributed among journals?



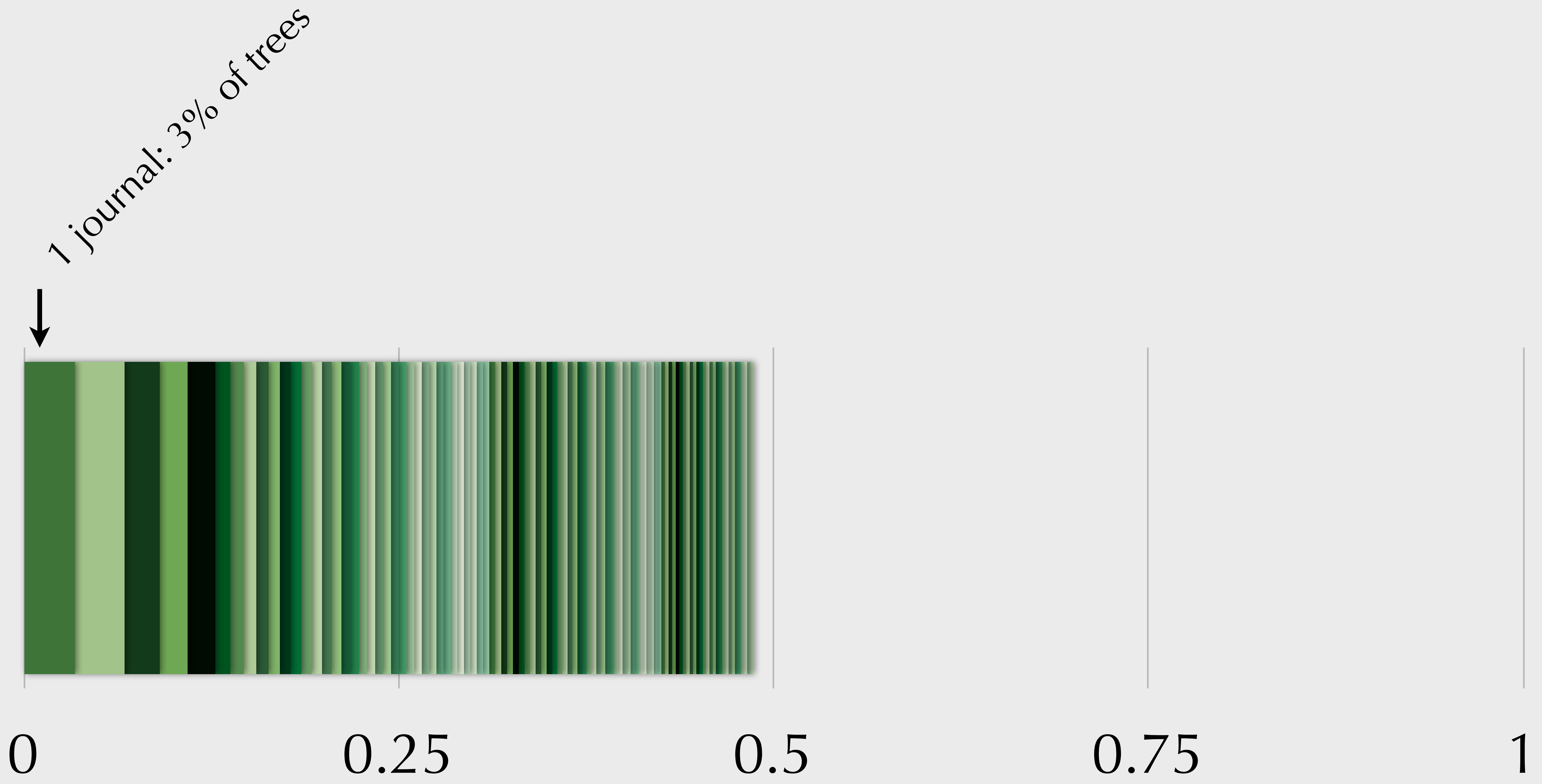
Tuesday, November 1, 2011

The problem is that JDAP only covers less than a dozen journals, and these are not even the top dozen phylogeny-reporting journals. Out of all journals with phylogeny-related papers in 2010, . . .

click: The top journal only accounts for 3 % of hits. [Mol Phyl & Evol, which is a JDAP journal BTW]
click: The top 5 only account for 13 %
click, **click**: and so on
In other words, the publishing of trees is spread among a staggeringly large number of journals, the vast majority of which publish less than 20 trees in a year.

click: JDAP would have to expand its scope enormously to be effective.

How are phylogeny articles distributed among journals?



Proportion of published phylogenies

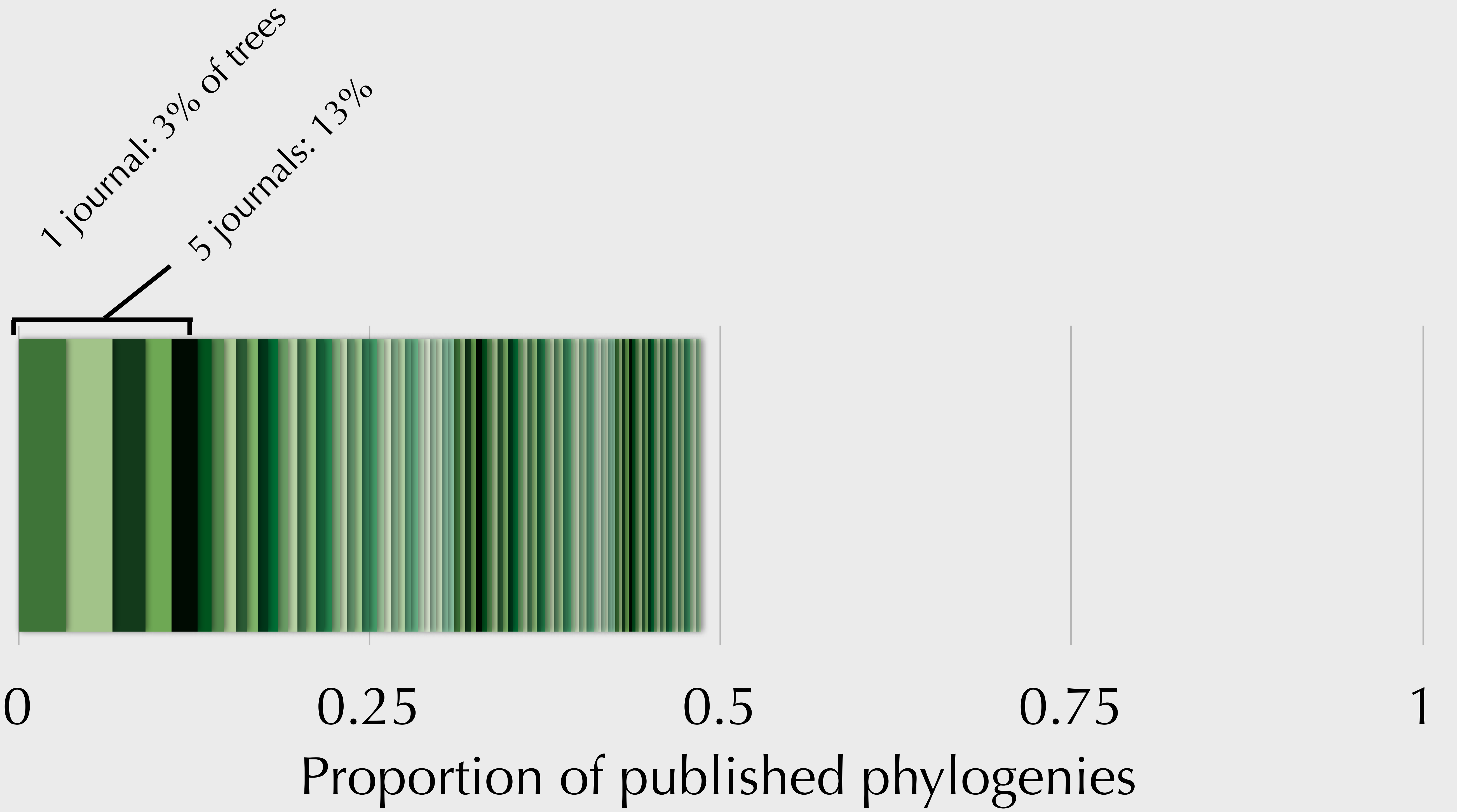
Tuesday, November 1, 2011

The problem is that JDAP only covers less than a dozen journals, and these are not even the top dozen phylogeny-reporting journals. Out of all journals with phylogeny-related papers in 2010, . . .

click: The top journal only accounts for 3 % of hits. [Mol Phyl & Evol, which is a JDAP journal BTW]
click: The top 5 only account for 13 %
click, **click**: and so on
In other words, the publishing of trees is spread among a staggeringly large number of journals, the vast majority of which publish less than 20 trees in a year.

click: JDAP would have to expand its scope enormously to be effective.

How are phylogeny articles distributed among journals?



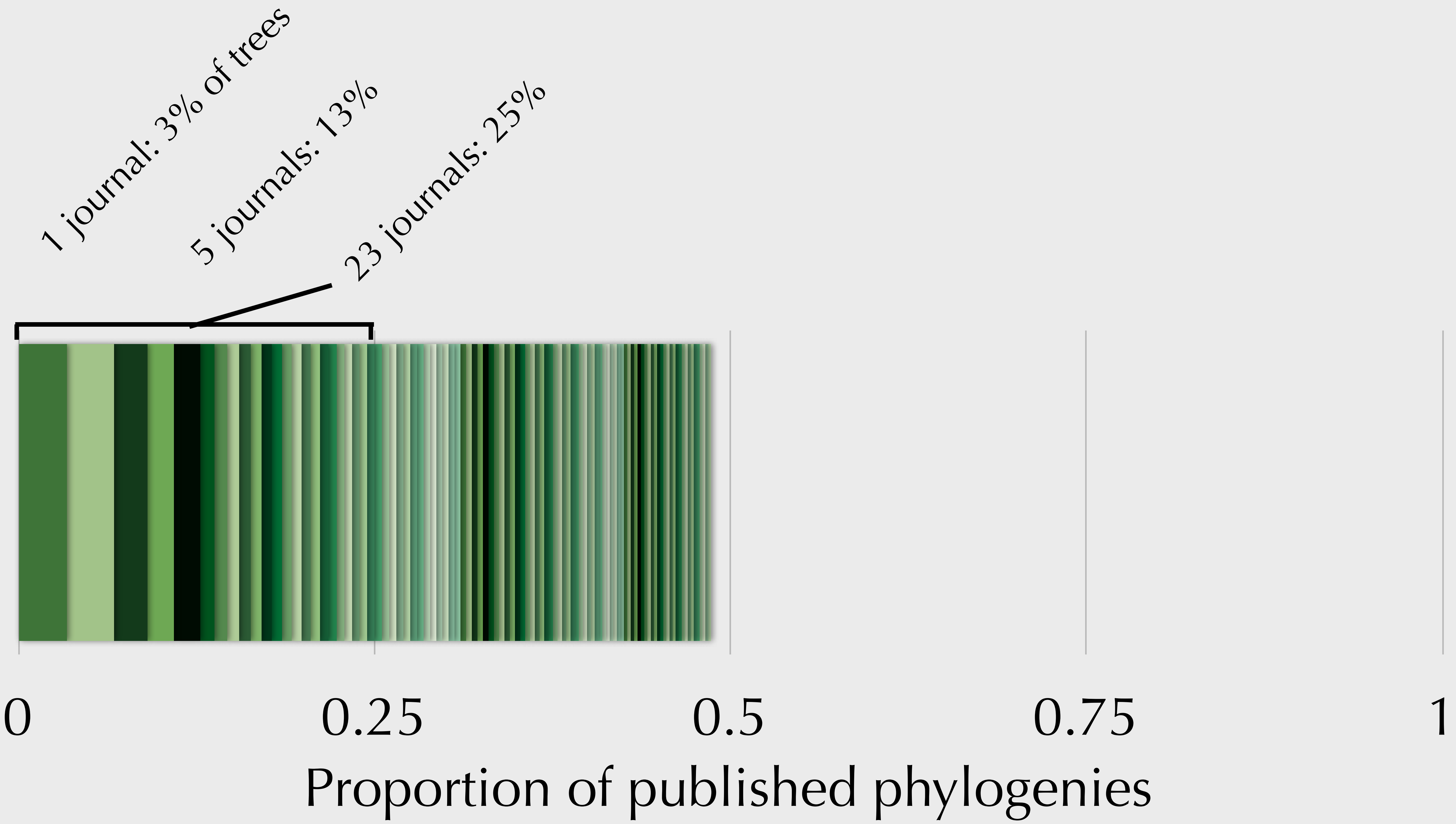
Tuesday, November 1, 2011

The problem is that JDAP only covers less than a dozen journals, and these are not even the top dozen phylogeny-reporting journals. Out of all journals with phylogeny-related papers in 2010, . . .

click: The top journal only accounts for 3 % of hits. [Mol Phyl & Evol, which is a JDAP journal BTW]
click: The top 5 only account for 13 %
click, **click**: and so on
In other words, the publishing of trees is spread among a staggeringly large number of journals, the vast majority of which publish less than 20 trees in a year.

click: JDAP would have to expand its scope enormously to be effective.

How are phylogeny articles distributed among journals?



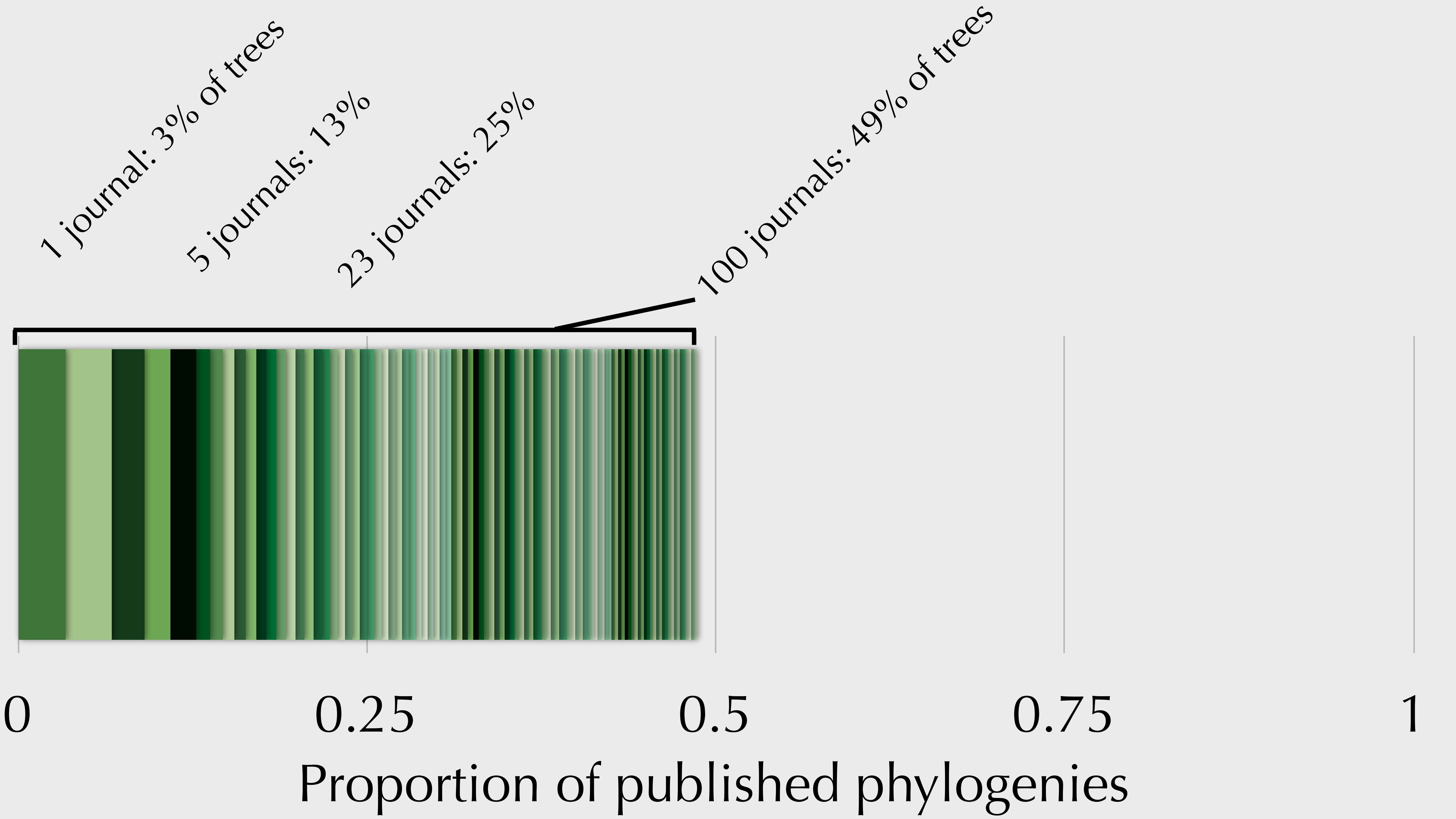
Tuesday, November 1, 2011

The problem is that JDAP only covers less than a dozen journals, and these are not even the top dozen phylogeny-reporting journals. Out of all journals with phylogeny-related papers in 2010, . . .

click: The top journal only accounts for 3 % of hits. [Mol Phyl & Evol, which is a JDAP journal BTW]
click: The top 5 only account for 13 %
click, **click**: and so on
In other words, the publishing of trees is spread among a staggeringly large number of journals, the vast majority of which publish less than 20 trees in a year.

click: JDAP would have to expand its scope enormously to be effective.

How are phylogeny articles distributed among journals?



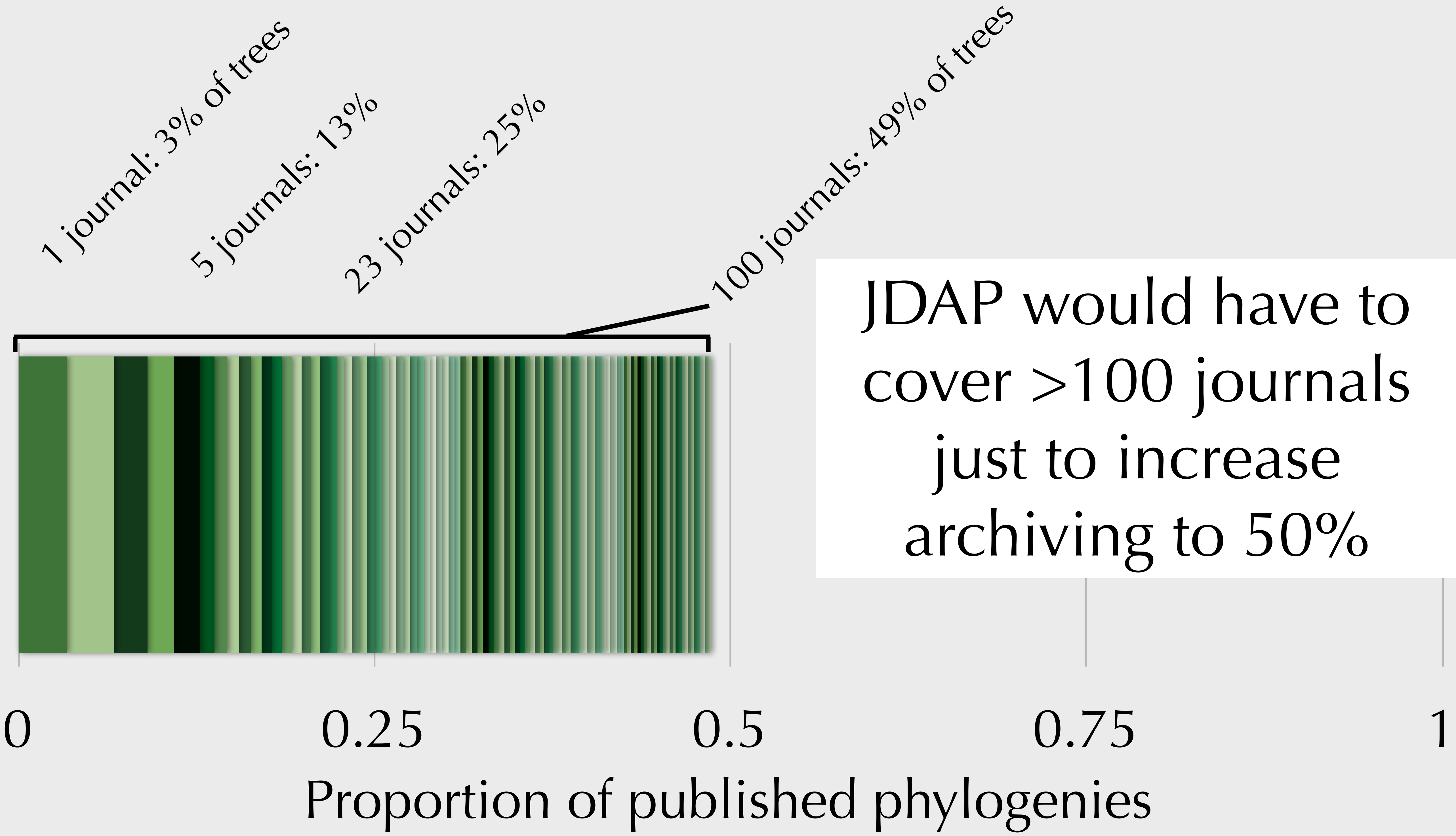
Tuesday, November 1, 2011

The problem is that JDAP only covers less than a dozen journals, and these are not even the top dozen phylogeny-reporting journals. Out of all journals with phylogeny-related papers in 2010, . . .

click: The top journal only accounts for 3 % of hits. [Mol Phyl & Evol, which is a JDAP journal BTW]
click: The top 5 only account for 13 %
click, **click**: and so on
In other words, the publishing of trees is spread among a staggeringly large number of journals, the vast majority of which publish less than 20 trees in a year.

click: JDAP would have to expand its scope enormously to be effective.

How are phylogeny articles distributed among journals?



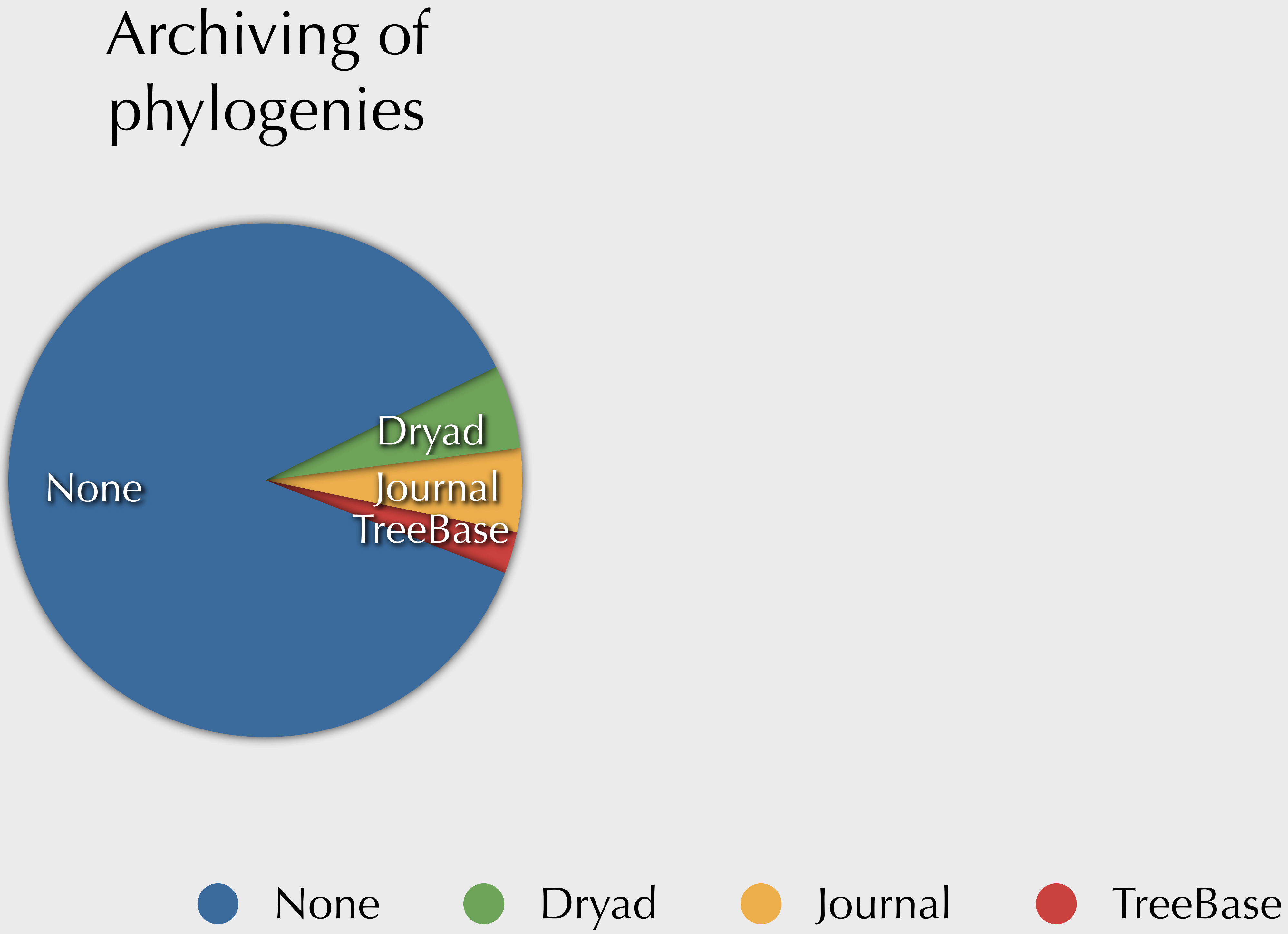
JDAP would have to cover >100 journals just to increase archiving to 50%

The problem is that JDAP only covers less than a dozen journals, and these are not even the top dozen phylogeny-reporting journals. Out of all journals with phylogeny-related papers in 2010, . . .

click: The top journal only accounts for 3 % of hits. [Mol Phyl & Evol, which is a JDAP journal BTW]
click: The top 5 only account for 13 %
click, **click**: and so on
In other words, the publishing of trees is spread among a staggeringly large number of journals, the vast majority of which publish less than 20 trees in a year.

click: JDAP would have to expand its scope enormously to be effective.

A funny thing we noticed

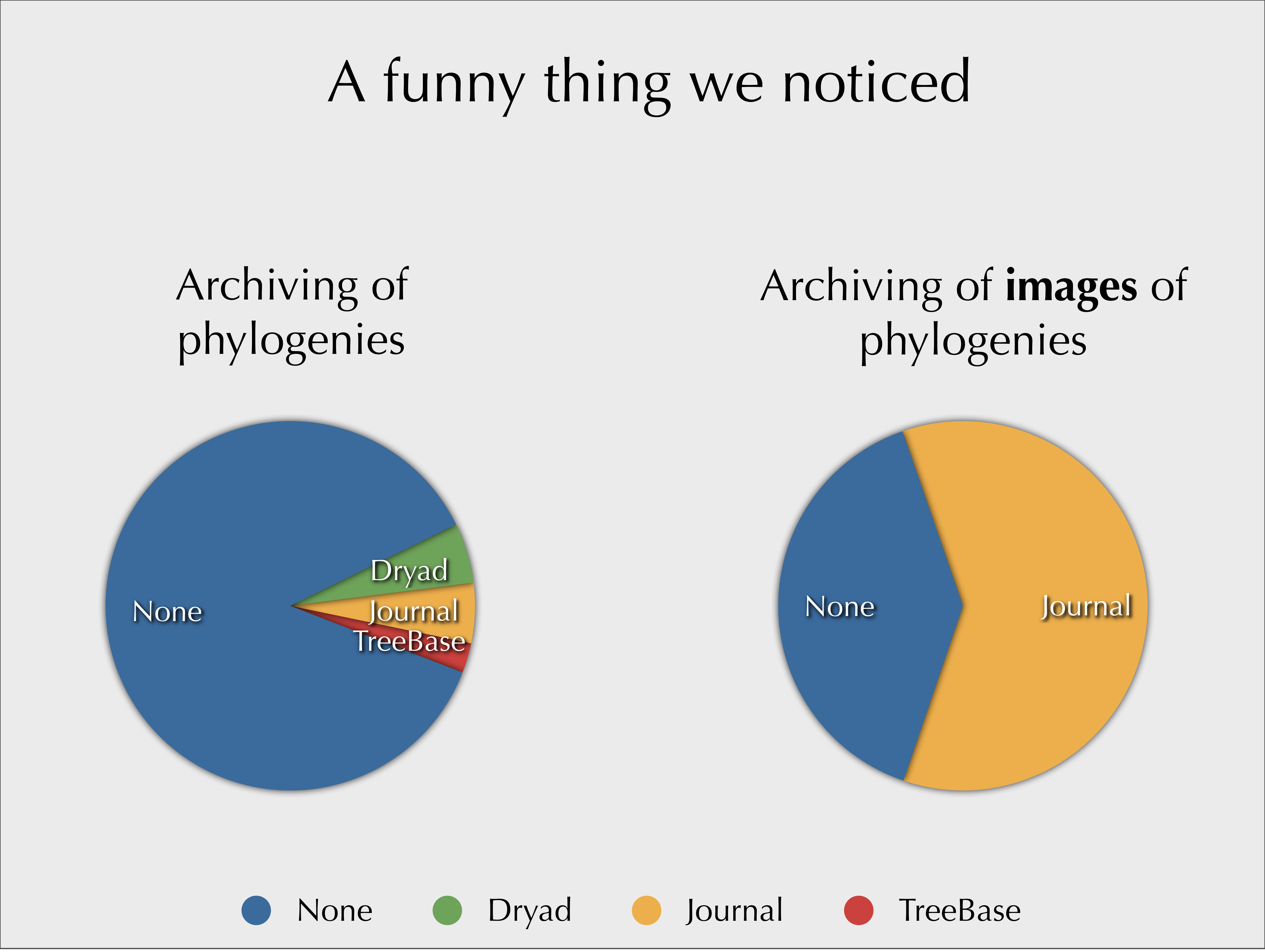


In May, 2011, we searched “phylogen” in Web of Sci and picked 40 recent papers from the top of a list sorted by “relevance” [what Web-o-Sci means by “relevance” is a trade secret, unfortunately]. Nearly all of these were published in the spring of 2011. We read each paper and looked for generation, re-use, or archiving of comparative data and trees.

For the 38 (out of 40) papers that generated a new phylogeny, 3 papers had archived results in Dryad and TreeBASE. In addition, authors sometimes provide trees in the “supplementary data” for their paper, stored on a journal web site.

Oddly, while few authors are archiving decodable versions of trees, most studies include images of extra trees in a supplementary data archive on the publisher’s web site. While these images may help the reader to evaluate the article, they do not represent an important pathway for re-use. (go quickly to next slide)

[The frequency of archiving in public archives in this small recent sample is 3 out of 38, as opposed to the 1.5 out of 38 we would expect from our large-scale analysis of 2010 publications.]



A funny thing we noticed

Archiving of phylogenies

Archiving of **images** of phylogenies

● None ● Dryad ● Journal ● TreeBase

In May, 2011, we searched “phylogen” in Web of Sci and picked 40 recent papers from the top of a list sorted by “relevance” [what Web-o-Sci means by “relevance” is a trade secret, unfortunately]. Nearly all of these were published in the spring of 2011. We read each paper and looked for generation, re-use, or archiving of comparative data and trees.

For the 38 (out of 40) papers that generated a new phylogeny, 3 papers had archived results in Dryad and TreeBASE. In addition, authors sometimes provide trees in the “supplementary data” for their paper, stored on a journal web site.

Oddly, while few authors are archiving decodable versions of trees, most studies include images of extra trees in a supplementary data archive on the publisher’s web site. While these images may help the reader to evaluate the article, they do not represent an important pathway for re-use. (go quickly to next slide)

[The frequency of archiving in public archives in this small recent sample is 3 out of 38, as opposed to the 1.5 out of 38 we would expect from our large-scale analysis of 2010 publications.]

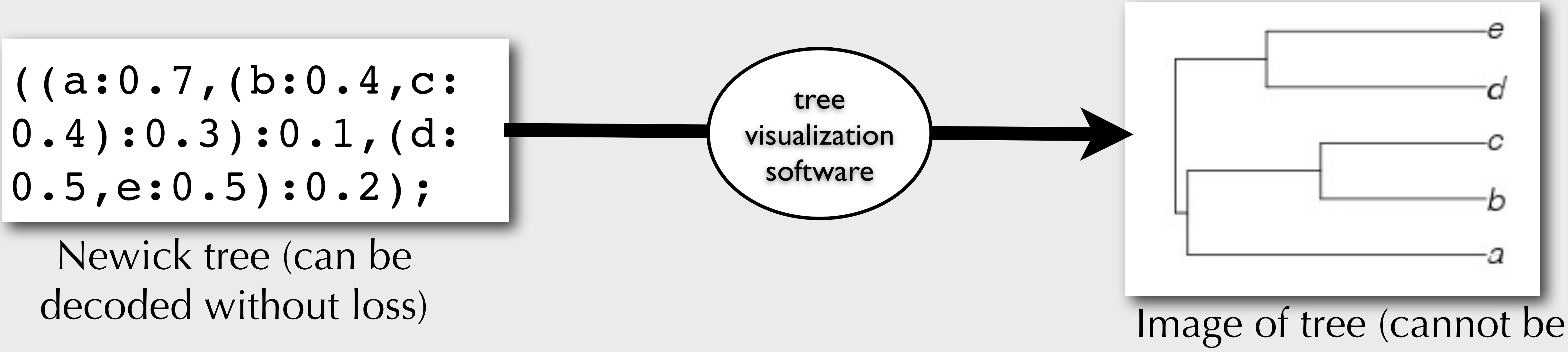
Thirsty?

Tuesday, November 1, 2011

Just in case the distinction isn't clear: this is not a glass of water. It is an image of a glass of water. There is a difference.

click: There is no technical barrier to saving a Newick treestring (left) in a text file on a journal web site. We know that the authors already have these treestrings, because they are using them to generate the tree images. As I will point out later, this suggests that one of the main barriers to sharing is a lack of awareness of when and how to make data share-able.

Thirsty?



```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

Newick tree (can be decoded without loss)

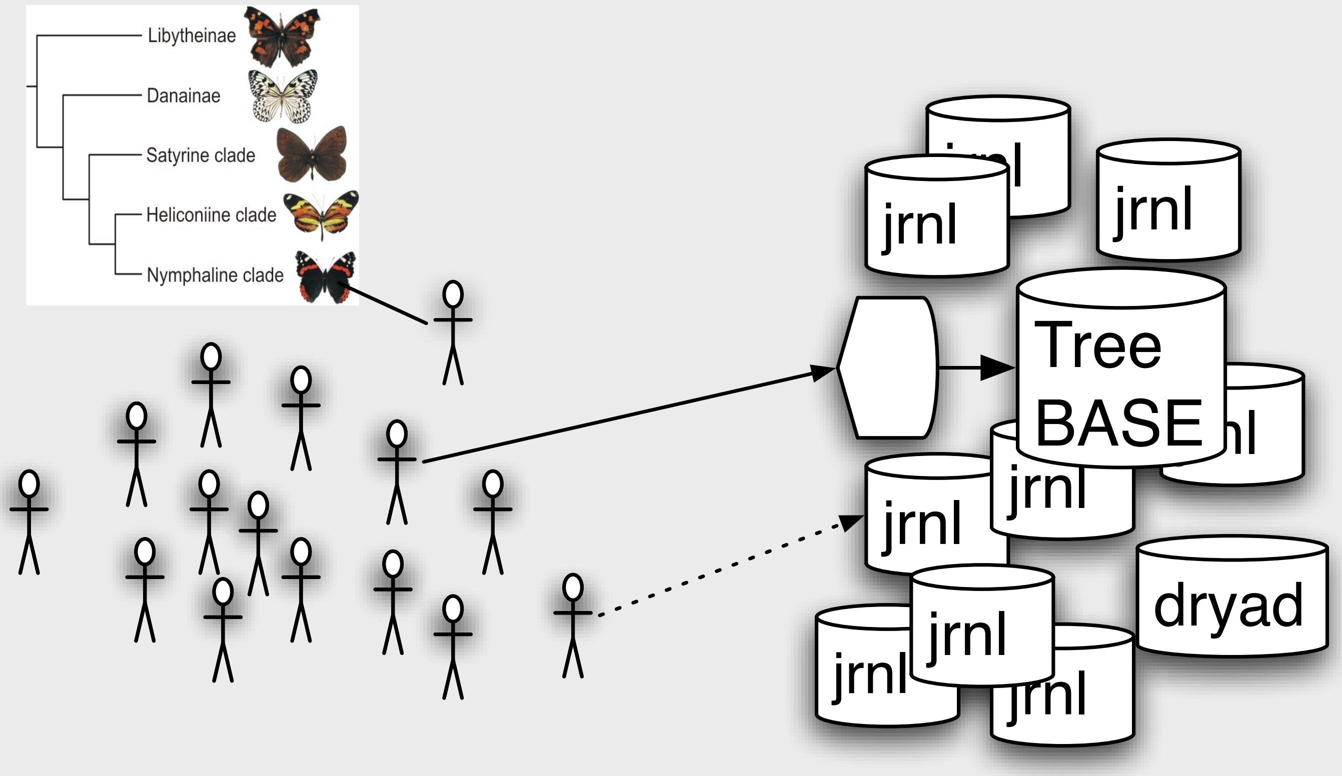
Image of tree (cannot be decoded without loss of information)

Tuesday, November 1, 2011

Just in case the distinction isn’t clear: this is not a glass of water. It is an image of a glass of water. There is a difference.

click: There is no technical barrier to saving a Newick treestring (left) in a text file on a journal web site. We know that the authors already have these treestrings, because they are using them to generate the tree images. As I will point out later, this suggests that one of the main barriers to sharing is a lack of awareness of when and how to make data share-able.

Sharing trees



tree image from nymphalidae.net

Producers

Consumers (re-users)

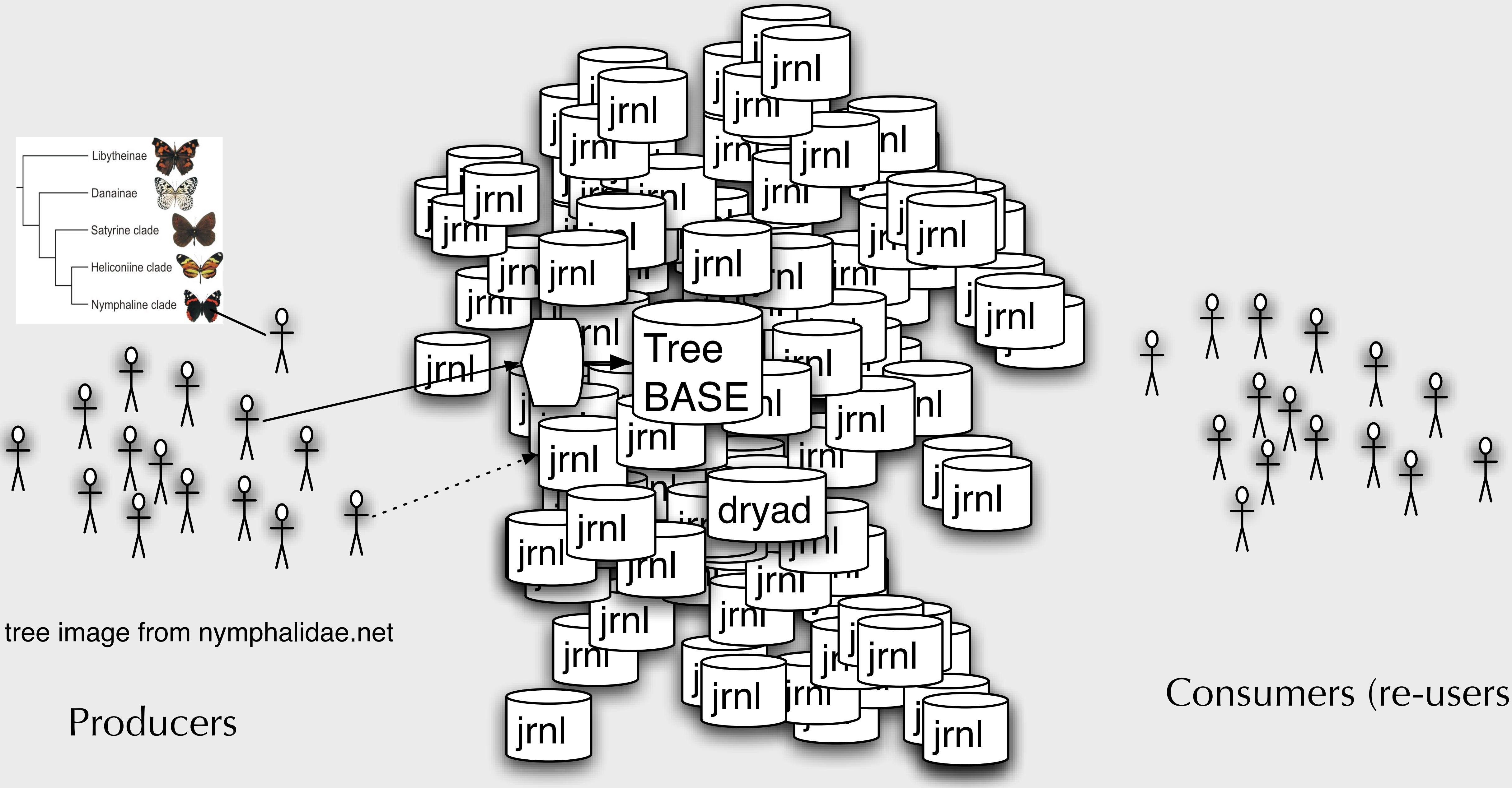
Repositories

Tuesday, November 1, 2011

(brief, just a few seconds) So, to summarize, producers sometimes archive trees in a re-usable way, but this is not common. The number of places where a tree could be archived is enormous.

To complete the picture, we need to understand how consumers are re-using these and other data.

Sharing trees



tree image from nymphalidae.net

Producers

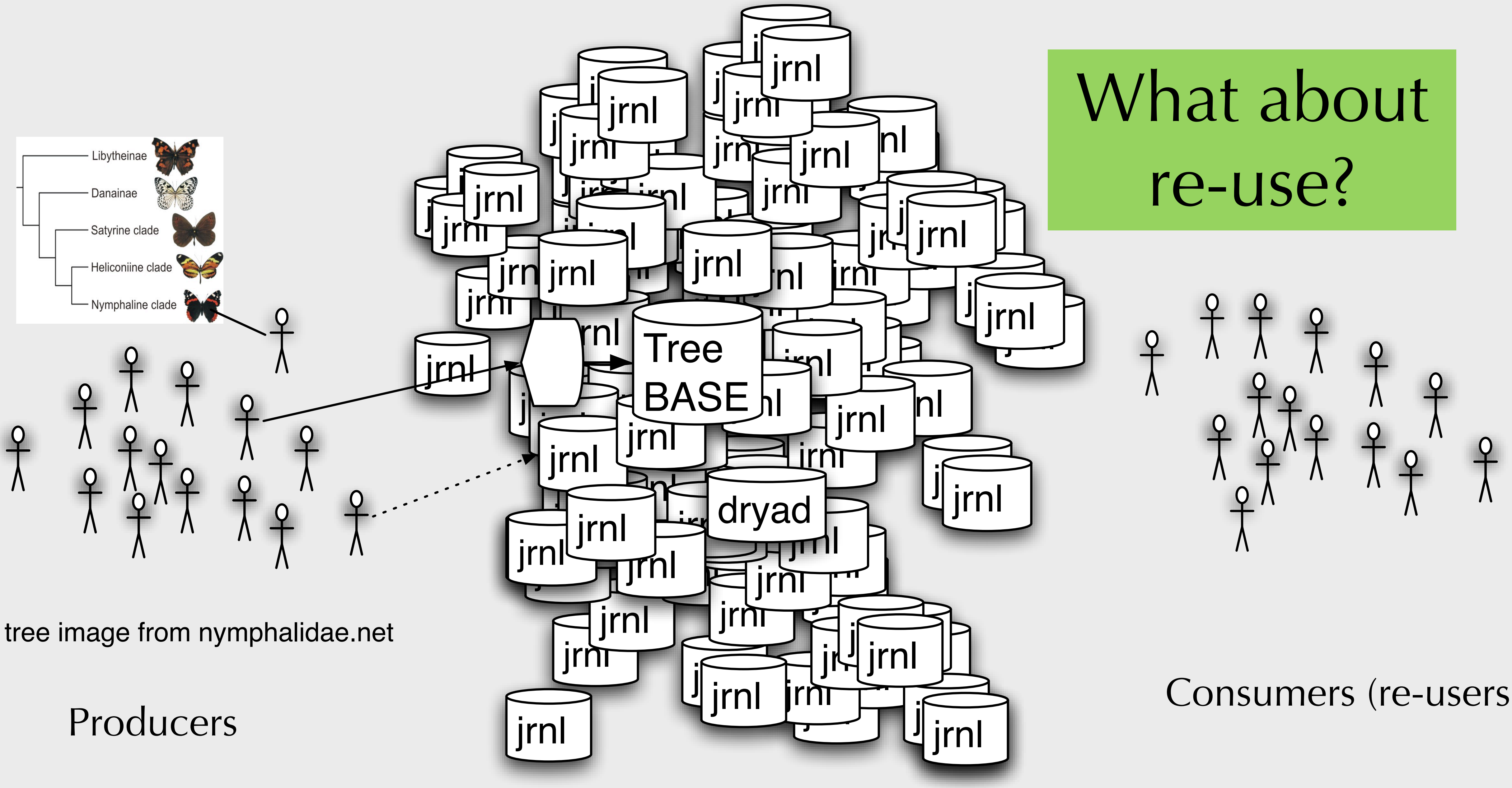
Consumers (re-users)

Repositories

(brief, just a few seconds) So, to summarize, producers sometimes archive trees in a re-usable way, but this is not common. The number of places where a tree could be archived is enormous.

To complete the picture, we need to understand how consumers are re-using these and other data.

Sharing trees



What about re-use?

tree image from nymphalidae.net

Producers

Consumers (re-users)

Repositories

(brief, just a few seconds) So, to summarize, producers sometimes archive trees in a re-usable way, but this is not common. The number of places where a tree could be archived is enormous.

To complete the picture, we need to understand how consumers are re-using these and other data.

To understand re-use, we have

- studied a sample of 40 recent articles (discovered via Web of Science)
- studied all articles in the April issues of Evolution and Am J. Bot
- interviewed users
- studied specific examples in more depth

Re-use of trees in a small sample of recent papers

Input tree	Research problem	Reference
APG tree	niche-diversity correlations	Burns & Strauss, 2011
APG tree	spatial distribution of wood traits	Zhang, et al., 2011
APG tree	spatial patterns of diversity	Morlon, et al., 2011
APG tree; Davies, et al tree	leaf veins & functions	Walls, 2011
Bininda-Emonds mammal tree	allometry of milk properties	Riek, 2011
APG tree	patch diversity	Duarte, 2010

Does anyone see a pattern here? Actually there are 3 patterns.

The first part of the pattern is that all of the cases of tree re-use in our study involved large species trees, not small species trees with fewer than 100 species, and not gene trees with paralogy.   Whatever one wants to call the input trees (I have heard different terms like “supertree” and “megatree” and I’m not sure what they mean) they are larger than the vast majority of trees— far larger than what most users would be willing to attempt. These are high-value trees, partly because generating them requires a special computational effort.

Researchers typically do not use these trees in whole form. Instead, they perform grafting and pruning operations, which is what Phylomatic was designed to do. The tree does not have all species, but if you know the genus name, the species can be attached to the genus node. This is the simplest possible grafting operation. “Pruning” means cutting away the species you don’t need. For instance, Ramona Walls, a post-doc at the NY botanical garden, had a set of data on leaf vein patterns for several hundred species, and she got a pruned Phylomatic tree with just those species, then used it in her analysis.

The second part of the pattern is that 5 of them use a super-tree for plants that comes from the plant tree-of-life community. One of them uses a mammal tree with over 4500 species.

A third pattern, not as obvious as the first two, is that these large species trees are being used for just two types of studies: studies of community ecology in which a phylogeny is used solely to measure diversity (Burns & Straus, Duarte, Morlon, Zhang), and functional trait analysis (Riek, Walls).

sources of the re-used trees

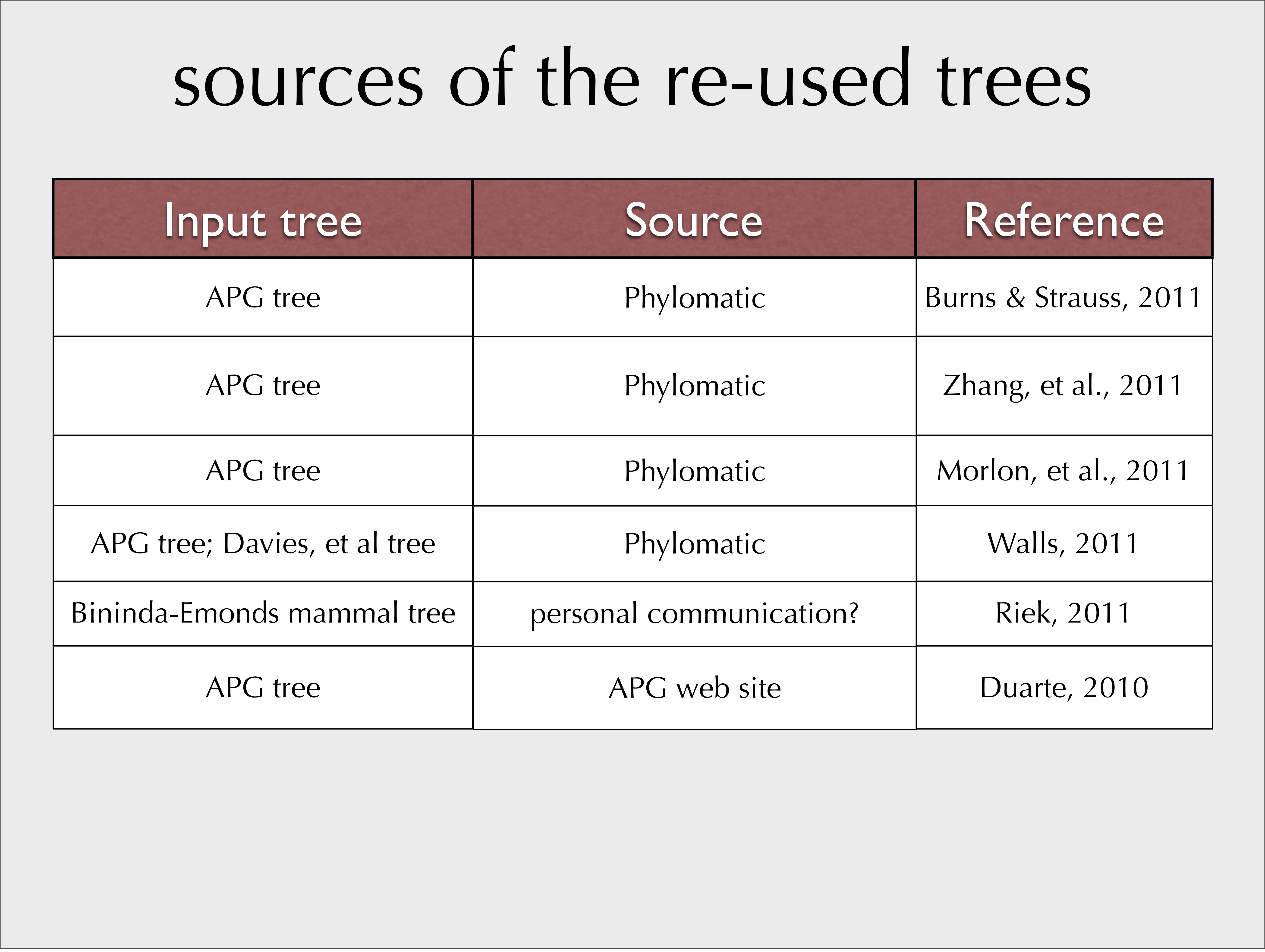
Input tree	Source	Reference
APG tree	?	Burns & Strauss, 2011
APG tree	?	Zhang, et al., 2011
APG tree	?	Morlon, et al., 2011
APG tree; Davies, et al tree	?	Walls, 2011
Bininda-Emonds mammal tree	?	Riek, 2011
APG tree	?	Duarte, 2010

Now, in order to construct our picture of data-sharing as information transfer, we need to know where users obtain these trees. Do they get them from archives such as TreeBASE and Dryad? No. From journal web sites? No.

Instead, they get them in other ways. In particular, users get these trees from Phylomatic. They also get the from project web sites and from personal communications.

What is Phylomatic?? It is not a public archive like TreeBASE or Dryad. It emerged to serve the plant tree-of-life community by providing an interface for practical re-use of species trees. Obviously, it's working, in spite of the fact that it is not a web-based program, but something that must be downloaded. In terms of serving the needs of scientists, Phylomatic rocks, while other tree-of-life projects whose names will not be mentioned, do not rock.

Remember that this is just from a set of 40 recent papers. We have looked at other small sets of papers, e.g., we looked at all articles in the April issues of Evolution and American Journal of Botany. In such arbitrary samples, we have not found any studies that are based on trees from TreeBASE or Dryad. However, we have seen several cases of trees that were used from a previous study. Often, but not always, the authors of the papers overlap, i.e., people are re-using their own previously published trees. It is rare for the whole tree to be re-used-- usually there is some grafting or pruning.



sources of the re-used trees

Input tree	Source	Reference
APG tree	Phylomatic	Burns & Strauss, 2011
APG tree	Phylomatic	Zhang, et al., 2011
APG tree	Phylomatic	Morlon, et al., 2011
APG tree; Davies, et al tree	Phylomatic	Walls, 2011
Bininda-Emonds mammal tree	personal communication?	Riek, 2011
APG tree	APG web site	Duarte, 2010

Tuesday, November 1, 2011

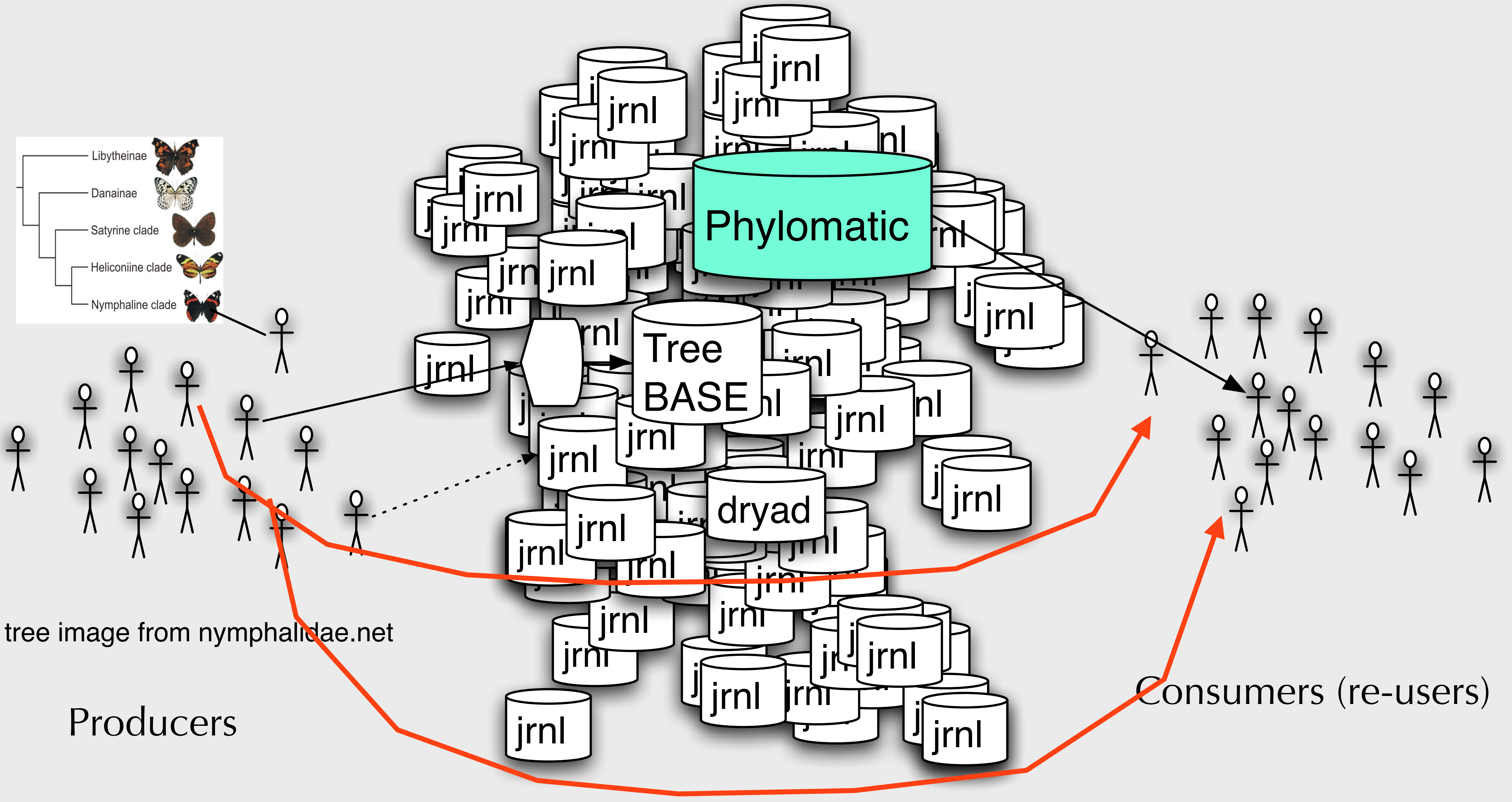
Now, in order to construct our picture of data-sharing as information transfer, we need to know where users obtain these trees. Do they get them from archives such as TreeBASE and Dryad? No. From journal web sites? No.

Instead, they get them in other ways. In particular, users get these trees from Phylomatic. They also get the from project web sites and from personal communications.

What is Phylomatic?? It is not a public archive like TreeBASE or Dryad. It emerged to serve the plant tree-of-life community by providing an interface for practical re-use of species trees. Obviously, it's working, in spite of the fact that it is not a web-based program, but something that must be downloaded. In terms of serving the needs of scientists, Phylomatic rocks, while other tree-of-life projects whose names will not be mentioned, do not rock.

Remember that this is just from a set of 40 recent papers. We have looked at other small sets of papers, e.g., we looked at all articles in the April issues of Evolution and American Journal of Botany. In such arbitrary samples, we have not found any studies that are based on trees from TreeBASE or Dryad. However, we have seen several cases of trees that were used from a previous study. Often, but not always, the authors of the papers overlap, i.e., people are re-using their own previously published trees. It is rare for the whole tree to be re-used-- usually there is some grafting or pruning.

Sharing trees



tree image from nymphalidae.net

Producers

Consumers (re-users)

Repositories

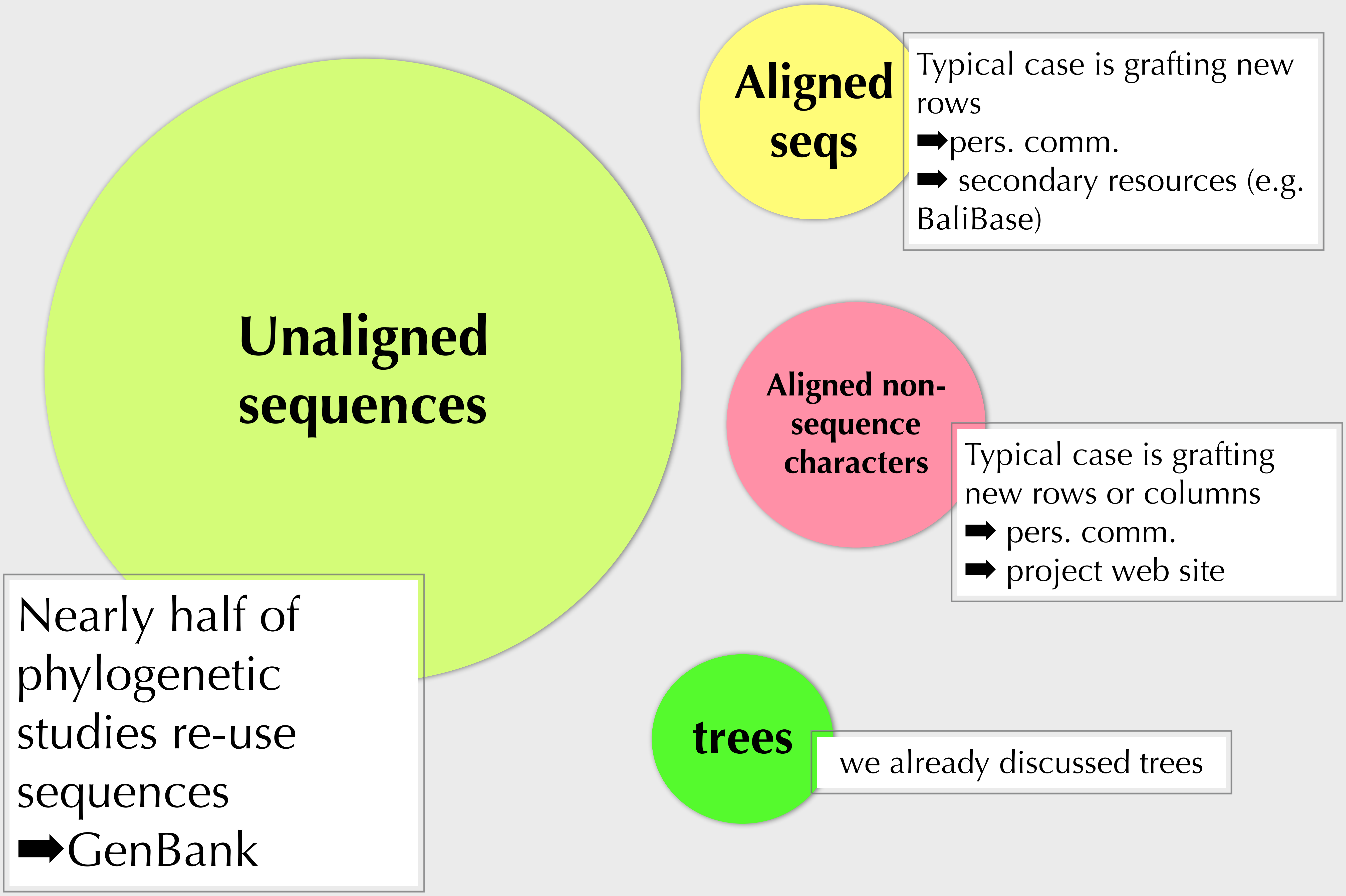
We can now summarize the big picture of sharing phylogenetic trees-- bearing in mind that this is based on a limited sample.

There are an enormous number of places for users to archive their trees. A small percentage of users, roughly 4%, archive their trees in a public archive. The rest of the trees have not been archived, or are spread out on various journal web sites.

Nevertheless, re-use of phylogenies is significant. Typically, when re-use happens, it happens in one of two ways: by phylomatic, or by personal communication.

[I haven’t drawn an input arrow for Phylomatic, because, even after reading the paper and checking out the project web site, it is not precisely clear where the phylogenies come from, or how they are made.]

what about other kinds of data?



Aligned seqs

Typical case is grafting new rows
➡pers. comm.
➡ secondary resources (e.g. BaliBase)

Unaligned sequences

Nearly half of phylogenetic studies re-use sequences
➡GenBank

Aligned non-sequence characters

Typical case is grafting new rows or columns
➡ pers. comm.
➡ project web site

trees

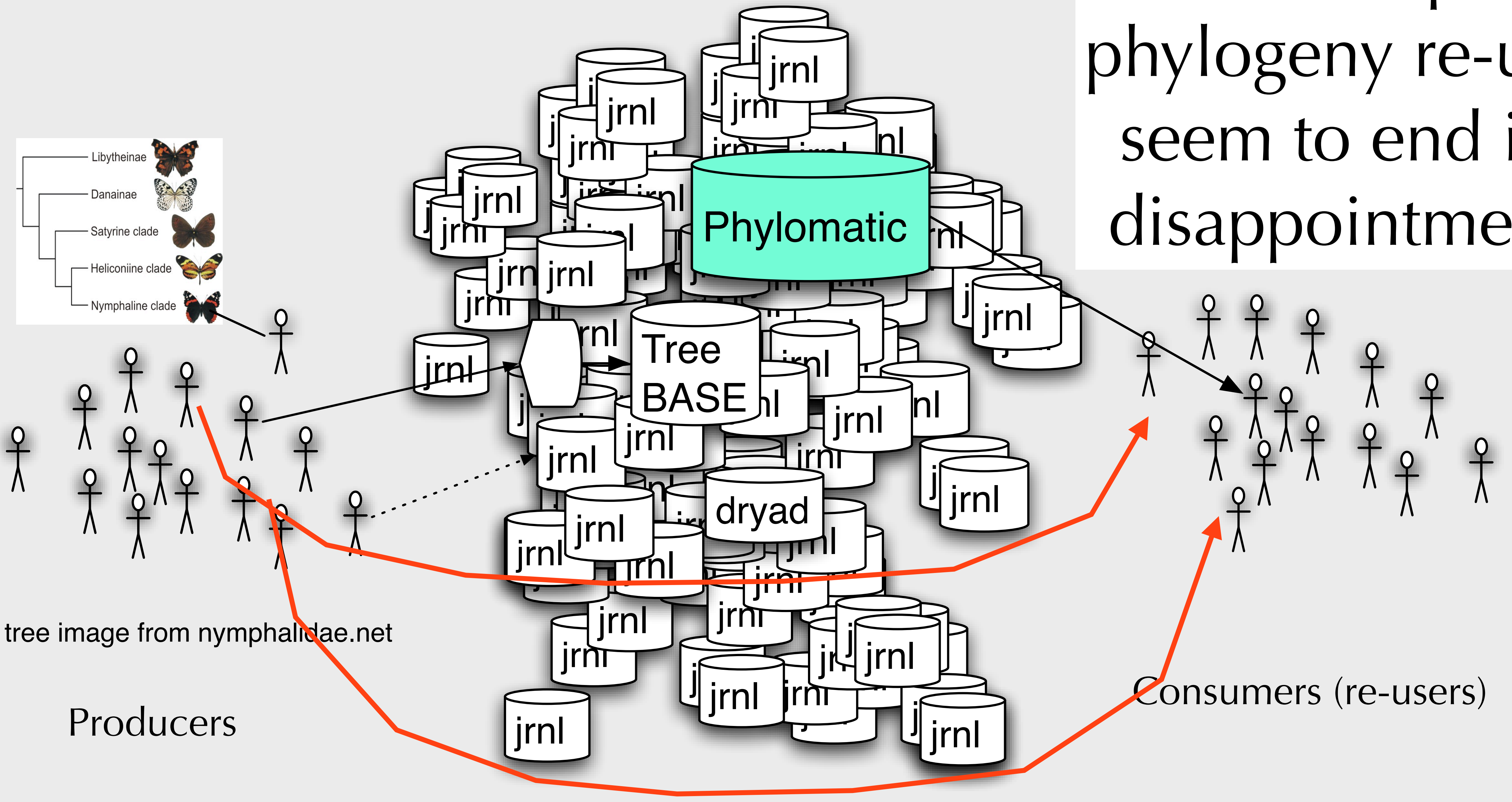
we already discussed trees

Tuesday, November 1, 2011

I'm focusing today on trees, but we also see frequent re-use of sequences, and less frequent re-use of aligned characters, including sequence characters as well as non-sequence characters such as morphological, behavioral, and physiological traits. The sequences come from GenBank, and the other kinds of data typically are shared via personal communications or project web sites.

Barriers to re-use

Most attempts at phylogeny re-use seem to end in disappointment



tree image from nymphalidae.net

Producers

Consumers (re-users)

Repositories

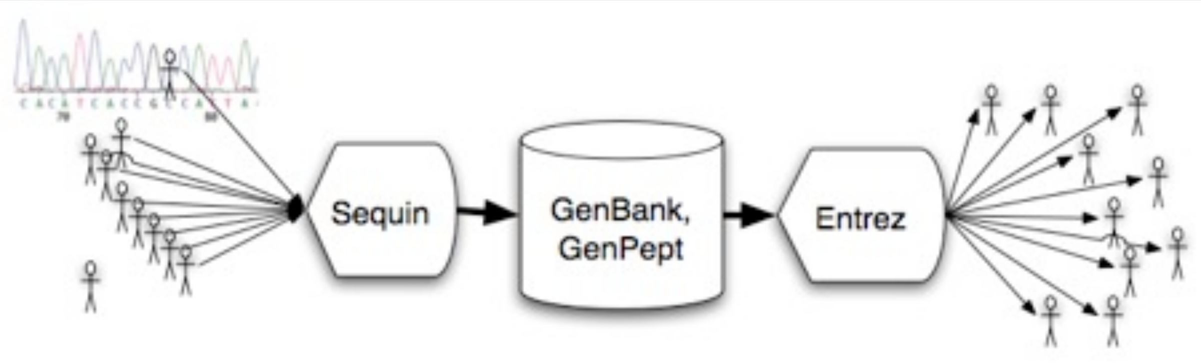
We have considered barriers to re-use by talking to users. Most attempts at re-use seem to end in disappointment.

By contrast, sharing of DNA sequences is highly successful. So, another way to study barriers is to compare the environment for sharing of sequences with that for sharing trees.

key barriers

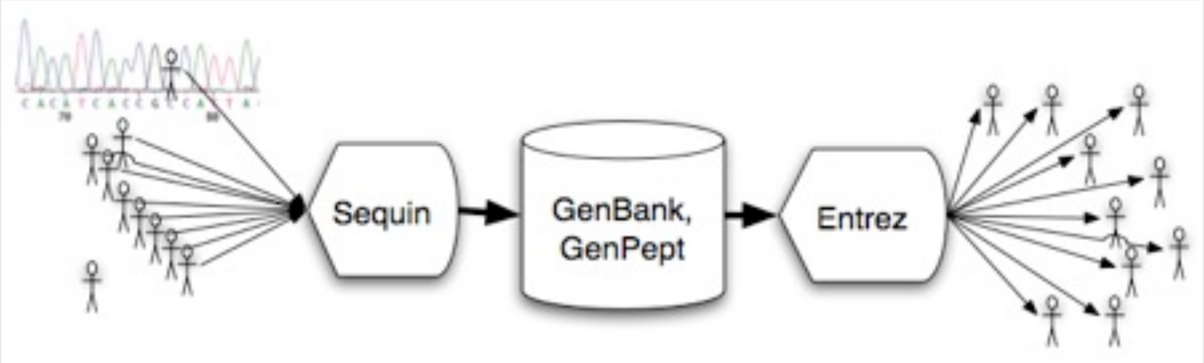
- discoverability: can’t find it
- access: located it, but can’t obtain it
- decoding: got it, but can’t decode it
- semantics: decoded it, but can’t interpret it or link it to other knowledge
- quality evaluation: I know what it means, but I’m not sure I can trust it

What enhances discoverability?



- centralization or aggregation - all records are discoverable via one interface
- standardized rich metadata - all records have rich metadata that a search interface can use
- critical links - records can be linked to other types of records, particularly publications

What enhances access?

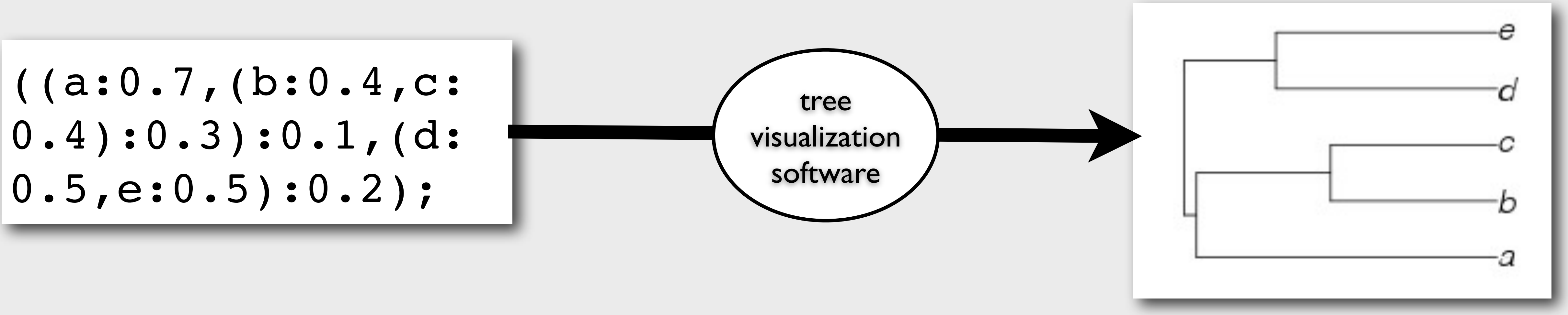


- public archiving - takes sharing out of the hands of individuals (who often refuse to share, or delay), removes paywalls

poor access	improved access
"Available from the author upon request"	"Available from Dryad Digital Repository doi:10.5061/dryad.34984"

What enhances decoding?

- use text-based formats - not images
- use data formats for data - e.g., CSV not PDF



```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

tree visualization software

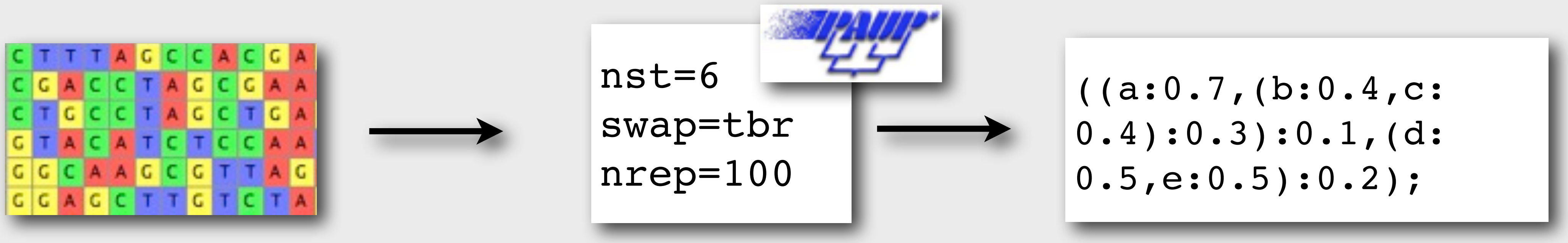
What enhances semantic integration?

- consistent, externally meaningful identifiers for biological entities
- annotations based on controlled vocabularies

hard to integrate	easy to integrate
((**Dmel**:0.7,(b:0.4,c:0.4):0.3)	((‘***Drosophila melanogaster*** **Meigen 1830**’:0.7,(b:0.4,c:0.4):0.3)

What enhances quality evaluation?

- methods annotations
- some system of third-party reviewing

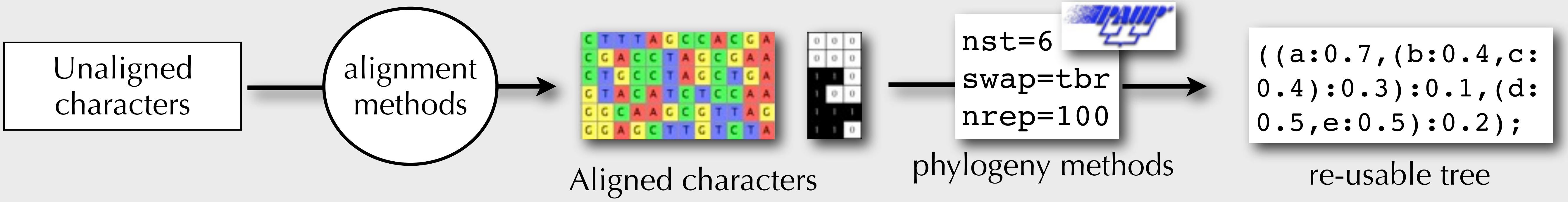


```
CTTTAGCCACGA
CGACCTAGCGAA
CTGCCTAGCTGA
GTACATCTCCAA
GGCAAGCGTTAG
GGAGCTTGTCTA
```

```
nst=6
swap=tbr
nrep=100
```

```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

A typical phylogenetics workflow



Unaligned characters

```
CTTTAGCCACGA
CGACCTAGCGAA
CTGCCTAGCTGA
GTACATCTCCAA
GGCAAGCGTTAG
GGAGCTTGTCTA
```

```
000
000
110
100
111
110
```

Aligned characters

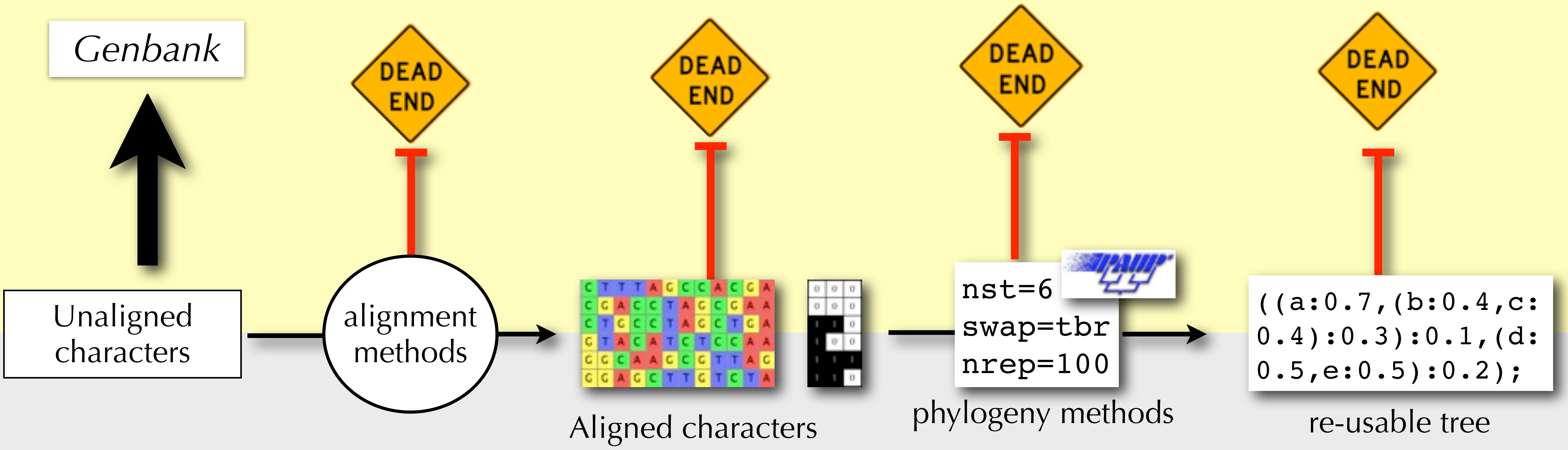
```
nst=6
swap=tbr
nrep=100
```

phylogeny methods

```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

re-usable tree

Current archiving practices (typical)



Genbank

DEAD END

DEAD END

DEAD END

DEAD END

Unaligned characters

alignment methods

```
C T T T A G C C A C G A
C G A C C T A G C G A A
C T G C C T A G C T G A
G T A C A T C T C C A A
G G C A A G C G T T A G
G G A G C T T G T C T A
```

```
0 0 0
0 0 0
1 1 0
1 0 0
1 1 1
1 1 0
```

Aligned characters

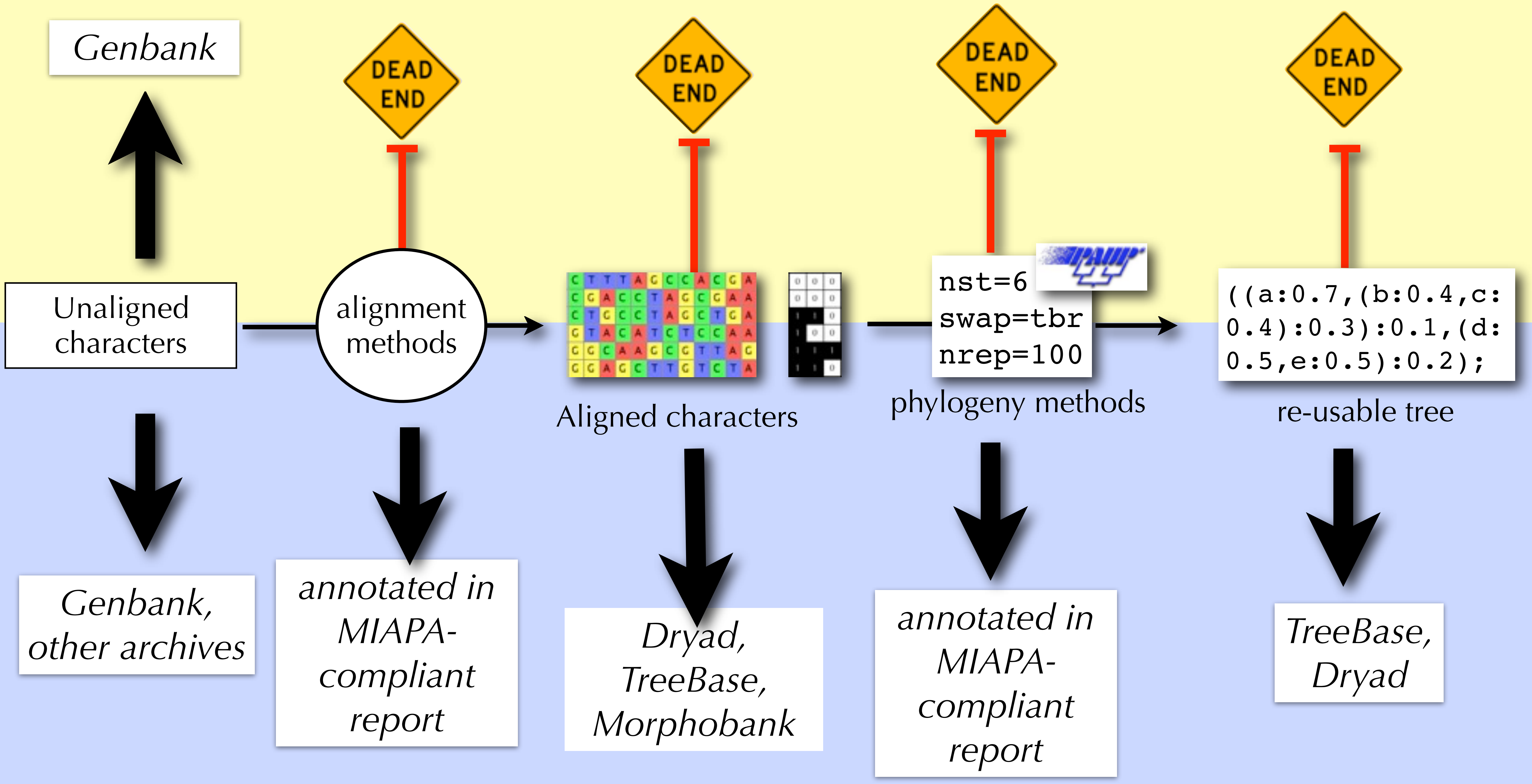
```
nst=6
swap=tbr
nrep=100
```

phylogeny methods

```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

re-usable tree

Current archiving practices (typical)



Genbank

Unaligned characters

alignment methods

Aligned characters

```
nst=6
swap=tbr
nrep=100
```

phylogeny methods

```
((a:0.7,(b:0.4,c:
0.4):0.3):0.1,(d:
0.5,e:0.5):0.2);
```

re-usable tree

Genbank, other archives

annotated in MIAPA-compliant report

Dryad, TreeBase, Morphobank

annotated in MIAPA-compliant report

TreeBase, Dryad

Future practices?

Arlin

Brian

Jamie

Ross

Dan

Thanks

- Bill Piel, Ryan Scherle, Todd Vision, Ryan Norris, Ramona Walls, Jeremy Wright, Heather Piwowar, others
- **MIAPA survey group** (Sudhir Kumar, Ross Mounce, Rutger Vos, Emily Gillespie, Nico Cellinese, Enrico Pontelli, Arlin Stoltzfus)
- TDWG and the **TDWG Phylogenetic Standards Interest**

Finally, thanks.