De-centralized but global: Redesigning biodiversity data aggregation for improved engagement and impact

Nico Franz, Ed Gilbert & Beckett Sterner

Biodiversity Knowledge Integration Center – Arizona State University

Keynote Session – June 10, 2019

3rd Annual Digital Data Conference – Yale University, New Haven, CO

Structure of presentation

0. Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

Structure of presentation

0. Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

1. "New at ASU" (and more generally): **The NEON Biorepository Data Portal** – a Darwin Core-based ecological monitoring and forecasting resource.

Structure of presentation

0. Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

1. "New at ASU" (and more generally): **The NEON Biorepository Data Portal** – a Darwin Core-based ecological monitoring and forecasting resource.

2. Rethinking centralized biodiversity data aggregation: Diagnosis and **components**

of a de-centralized complement.

Structure of presentation

0. Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

1. "New at ASU" (and more generally): **The NEON Biorepository Data Portal** – a Darwin Core-based ecological monitoring and forecasting resource.

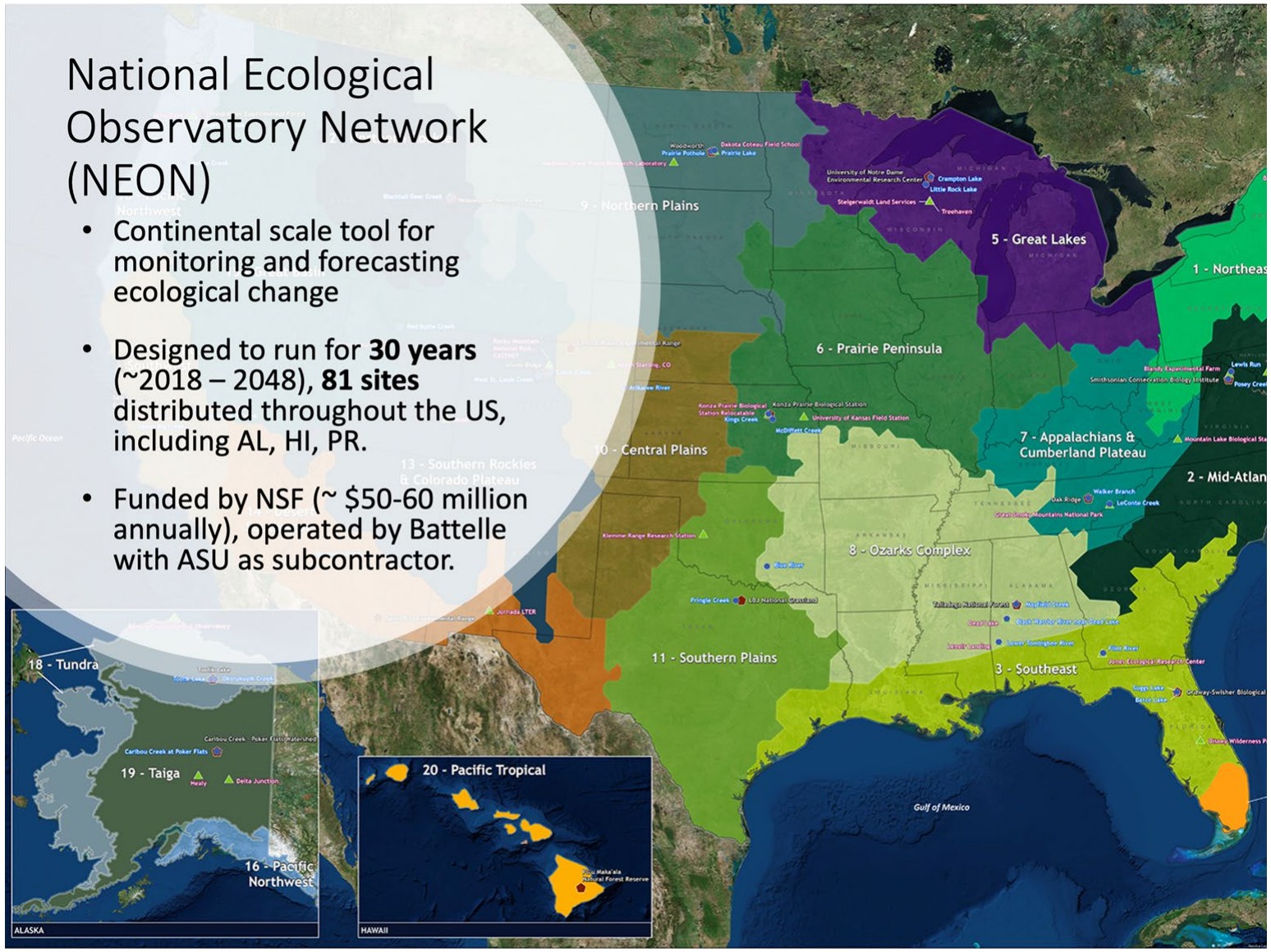
2. Rethinking centralized biodiversity data aggregation: Diagnosis and **components**

of a de-centralized complement.

Introducing the **NEON Biorepository Data Portal (1)**

https://biorepo.neonscience.org

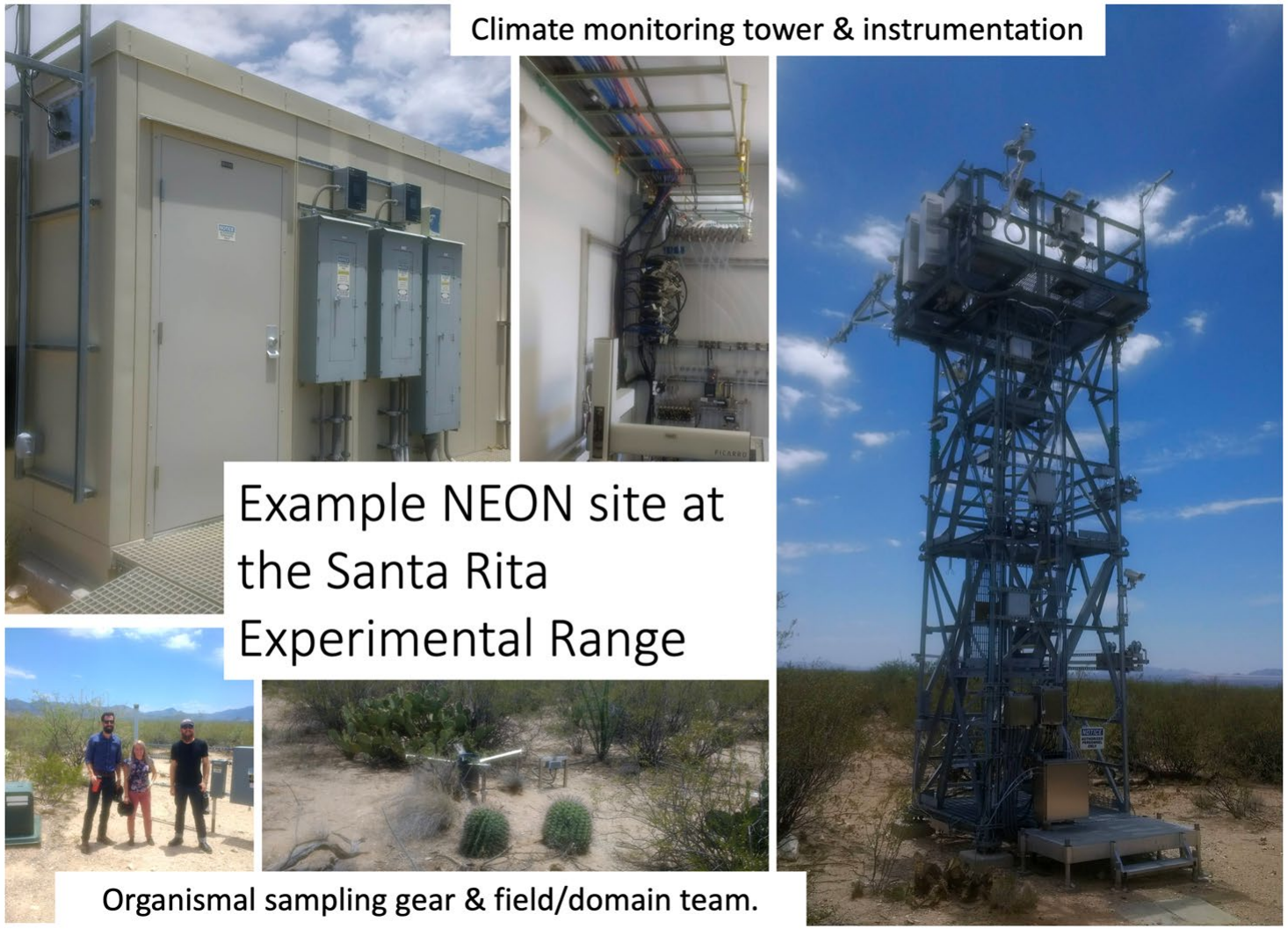
Introducing the **NEON Biorepository Data Portal**



Introducing the **NEON Biorepository Data Portal**

Nati
Obs
(NE

- Co
  mo
  ec
- De
  (~
  dis
  inc
- Fu
  an
  wi



Climate monitoring tower & instrumentation

Example NEON site at the Santa Rita Experimental Range

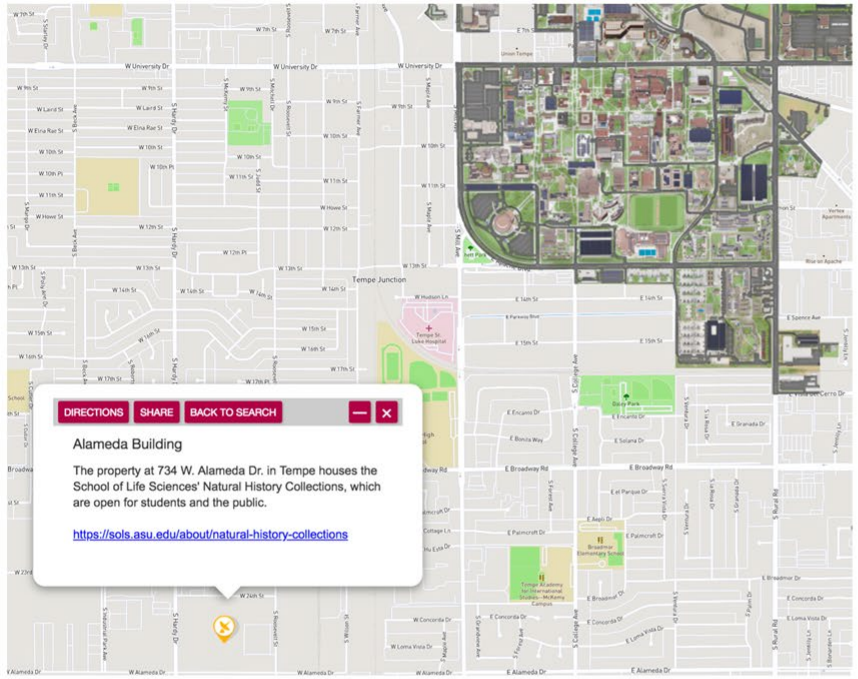
Organismal sampling gear & field/domain team.

Introducing the **NEON Biorepository Data Portal**

Domain facilities process samples and and send them to the NEON Biorepository at ASU.

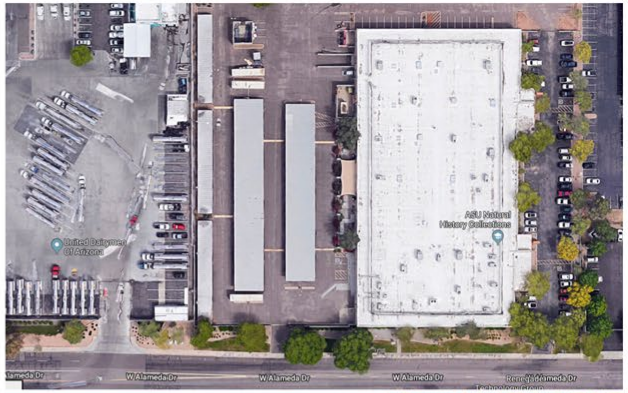
Introducing the **NEON Biorepository Data Portal**

Where are we located?



734 W Alameda Drive, Tempe, AZ 85282
https://www.asu.edu/map/interactive/?id=120&mrklid=66206

- Renovated and activated in the Fall of 2014.
- **Open to the public** (weekdays 9 am – 5 pm).
- **Free parking** available right at entrance.



Introducing the **NEON Biorepository Data Portal**

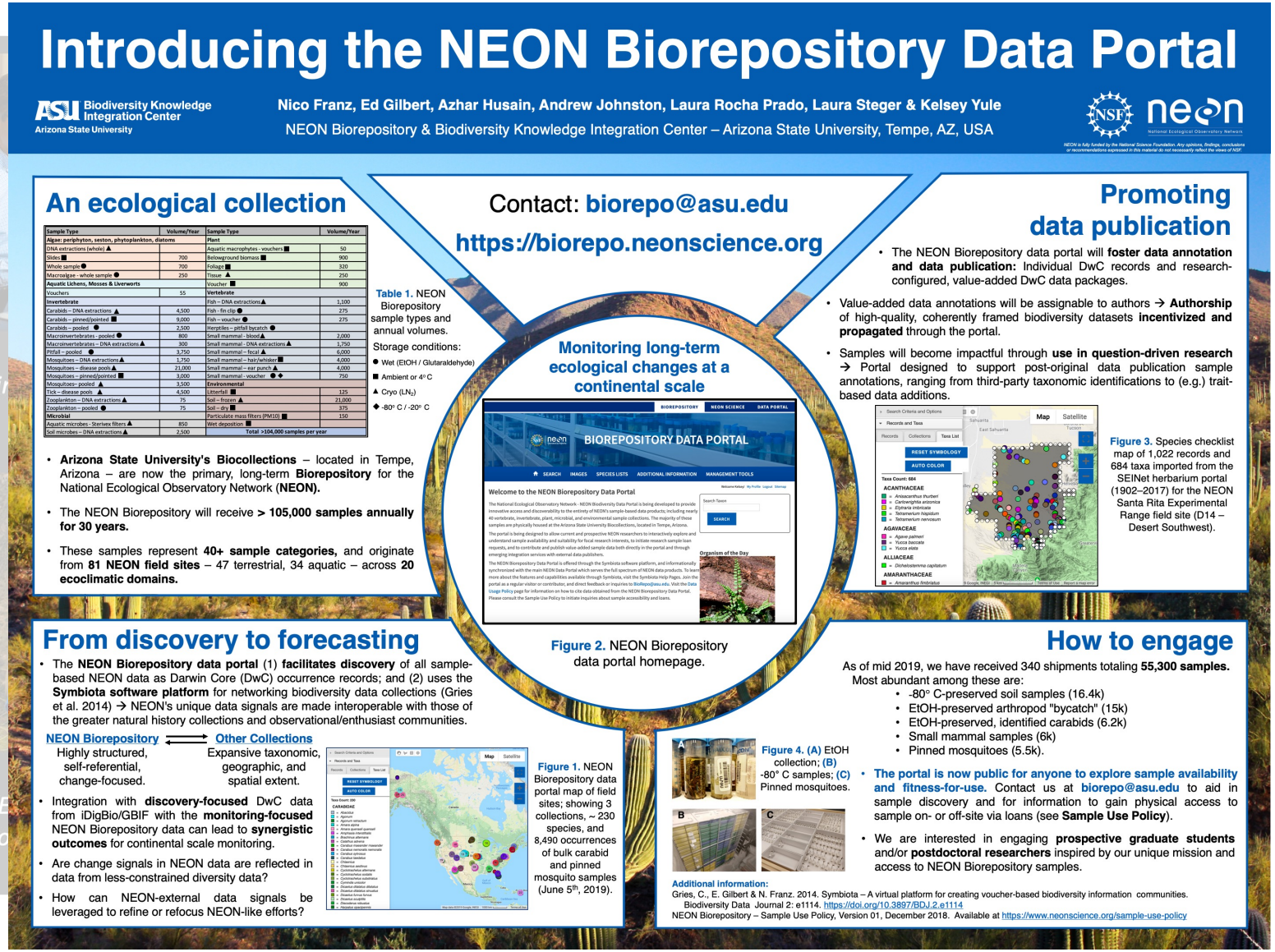
Whe

https://v

- Renov
- **Open**
- **Free p**

More information at Tuesday's poster session!

- Using the portal
- Access to samples
- New developments
  - Informatics
  - Research
- Prospective M.Sc./ Ph.D. applicants and postdocs
- Early and broad

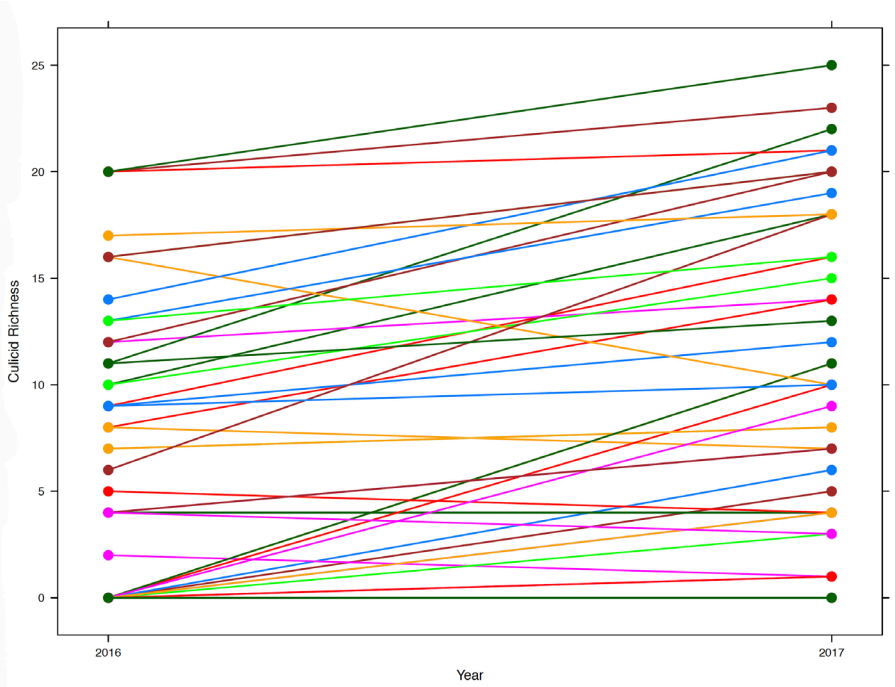


A short demonstration

https://biorepo.neonscience.org

[after a short BioRepo portal demo (Culicidae / SRER Plants)] *

Mosquito species richness per site, 2016–2017.



Change in species richness per site.

- NEON organismal sampling produces a unique, taxon-/region-/time-constrained, **change-focused data signal.**

* Thanks to Dr. Kelsey Yule, NEON Biorepository @ ASU.

Rethinking centralized biodiversity data aggregation (2)

OXFORD ACADEMIC

DATABASE
The Journal of Biological Databases and Curation

Volumes ▼ Submit ▼ Alerts About ▼ All Database

Volume 2018
2018

Article Contents

Abstract
Introduction
No simple diagnosis for data quality deficiencies
Importance of trust for sustained use
Matching accountability to responsibility
Generation of novel systematic syntheses
Role of the DwC standard
Aggregation and authorship
Consequences of disenfranchizing taxonomic experts
Backbone-based data signal distortion: an example
An objection
Conclusion
Acknowledgements
Funding

To increase trust, change the social design behind aggregated biodiversity data

Nico M Franz ✉, Beckett W Sterner

Database, Volume 2018, 2018, bax100, https://doi.org/10.1093/database/bax100
Published: 04 January 2018 **Article history** ▼

PDF Split View Cite Permissions

Abstract

Growing concerns about the quality of aggregated biodiversity data are lowering trust in large-scale data networks. Aggregators frequently respond to quality concerns by recommending that biologists work with original data providers to correct errors 'at the source.' We show that this strategy falls systematically short of a full diagnosis of the underlying causes of distrust. In particular, trust in an aggregator is not just a feature of the data signal quality provided by the sources to the aggregator, but also a consequence of the social design of the aggregation process and the resulting power balance between individual data contributors and aggregators. The latter have created an accountability gap by downplaying the authorship and significance of the taxonomic hierarchies—frequently called 'backbones'—they generate, and which are in effect novel classification theories that operate at the core of data-structuring process. The Darwin Core standard for sharing occurrence records plays an under-appreciated role in maintaining the accountability gap, because this standard lacks the syntactic structure needed to preserve the taxonomic coherence of data packages submitted for aggregation, potentially leading to inferences that no individual source would support. Since high-quality data packages can mirror competing and conflicting classifications, i.e. unsettled systematic research, this plurality must be accommodated in the design of biodiversity data integration. Looking forward, a key directive is to develop new technical pathways and social incentives for experts to contribute directly to the validation of taxonomically coherent data packages as part of a greater, trustworthy aggregation process.

Issue Section: Perspective/Opinion

- https://doi.org/10.1093/database/bax100
- Critique of monolithic aggregation designs.

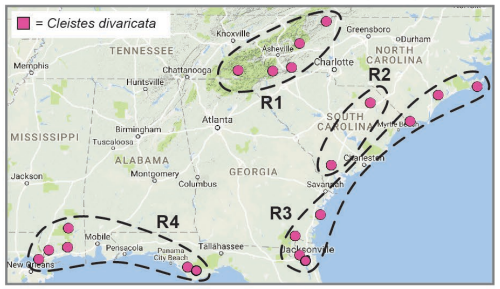
Alignment of 5 taxonomic schemata (1922 – present)

#	Period of use	Concept lineage 1	Concept lineage 2	Concept lineage 3	According to (sec.)
5	2009 – present	*Cleistesiopsis divaricata*	*Cleistesiopsis oricamporum*	*Cleistesiopsis bifaria*	Weakley (2015)
4	2008 – present	*Cleistesiopsis divaricata*	*Cleistesiopsis bifaria*		Kartesz (2010)
3	1993 – present	*Cleistes divaricata*	*Cleistes bifaria*		USDA Plants (2012)
2	1946 – 1993	*Cleistes div.* var. *divaricata*	*Cleistes divaricata* var. *bifaria*		Fernald (1950)
1	1922 – 1991	*Cleistes divaricata*			Radford AB (1968)

A. sec. Radford, Ahles & Bell (1968)

Distribution

	"*bifaria*"	"*divaric.*"	"*oricamp.*"
R1	o	+	o
R2	o	+	o
R3	o	+	o
R4	o	+	o

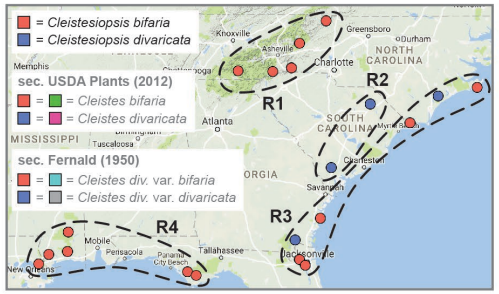


Data package 1
=> Ingestion

B. sec. Kartesz (2010) [BONAP]

Distribution

	bifaria	*divaric.*	*oricamp.*
R1	+	–	o
R2	–	+	o
R3	+	+	o
R4	+	–	o

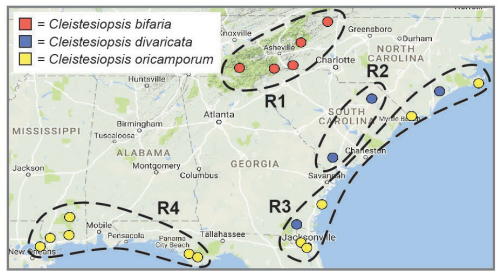


Data package 2
=> Ingestion

C. sec. Weakley (2015)

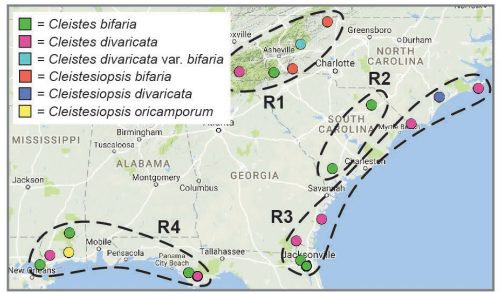
Distribution

	bifaria	*divaric.*	*oricamp.*
R1	+	–	–
R2	–	+	–
R3	–	+	+
R4	–	–	+



Data package 3
=> Ingestion

D. sec. SERNEC (2016) - RAW AGGREGATION



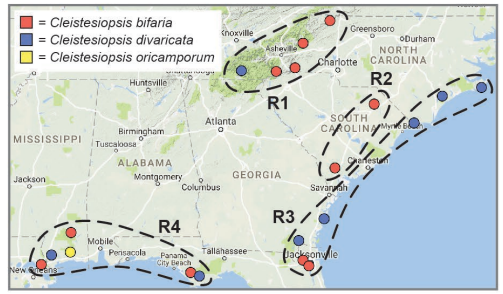
Distribution

	"*bifaria*"	"*divaric.*"	"*oricamp.*"
R1	⊕	⊕	–
R2	⊕	⊖	–
R3	+	+	–
R4	⊕	⊕	⊕

Data transformation to conform with single taxonomic backbone

Aggregation yields *novel inferences* of sympatry (R1,R4) & endemism (R2)

E. sec. SERNEC (2016) - AGGREGATOR SYNTHESIS



Distribution

	bifaria	*divaric.*	*oricamp.*
R1	⊕	⊕	–
R2	⊕	⊖	–
R3	+	+	–
R4	⊕	⊕	⊕

Rethinking centralized biodiversity data aggregation (2)

OXFORD ACADEMIC

DATABASE
The Journal of Biological Databases and Curation

Volumes ▼ Submit ▼ Alerts About ▼ All Database

Volume 2018
2018

Article Contents

Abstract
Introduction
No simple diagnosis for data quality deficiencies
Importance of trust for sustained use
Matching accountability to responsibility
Generation of novel systematic syntheses
Role of the DwC standard
Aggregation and authorship
Consequences of disenfranchizing taxonomic experts
Backbone-based data signal distortion: an example
An objection
Conclusion
Acknowledgements
Funding

To increase trust, change the social design behind aggregated biodiversity data

Nico M Franz ✉, Beckett W Sterner

Database, Volume 2018, 2018, bax100, https://doi.org/10.1093/database/bax100
Published: 04 January 2018 **Article history** ▼

PDF Split View Cite Permissions

Abstract

Growing concerns about the quality of aggregated biodiversity data are lowering trust in large-scale data networks. Aggregators frequently respond to quality concerns by recommending that biologists work with original data providers to correct errors 'at the source.' We show that this strategy falls systematically short of a full diagnosis of the underlying causes of distrust. In particular, trust in an aggregator is not just a feature of the data signal quality provided by the sources to the aggregator, but also a consequence of the social design of the aggregation process and the resulting power balance between individual data contributors and aggregators. The latter have created an accountability gap by downplaying the authorship and significance of the taxonomic hierarchies—frequently called 'backbones'—they generate, and which are in effect novel classification theories that operate at the core of data-structuring process. The Darwin Core standard for sharing occurrence records plays an under-appreciated role in maintaining the accountability gap, because this standard lacks the syntactic structure needed to preserve the taxonomic coherence of data packages submitted for aggregation, potentially leading to inferences that no individual source would support. Since high-quality data packages can mirror competing and conflicting classifications, i.e. unsettled systematic research, this plurality must be accommodated in the design of biodiversity data integration. Looking forward, a key directive is to develop new technical pathways and social incentives for experts to contribute directly to the validation of taxonomically coherent data packages as part of a greater, trustworthy aggregation process.

Issue Section: Perspective/Opinion

- https://doi.org/10.1093/database/bax100
- Critique of monolithic aggregation designs.
- Here we reformulate the issue in relation to the **direction of data flow** in hierarchical biodiversity data aggregation networks.

Rethinking centralized biodiversity data aggregation (2)

OXFORD ACADEMIC

DATABASE
The Journal of Biological Databases and Curation

Volumes ▼ Submit ▼ Alerts About ▼ All Database

DATABASE

Volume 2018
2018

Article Contents

Abstract
Introduction
No simple diagnosis for data quality deficiencies
Importance of trust for sustained use
Matching accountability to responsibility
Generation of novel systematic syntheses
Role of the DwC standard
Aggregation and authorship
Consequences of disenfranchizing taxonomic experts
Backbone-based data signal distortion: an example
An objection
Conclusion
Acknowledgements
Funding

To increase trust, change the social design behind aggregated biodiversity data

Nico M Franz ✉, Beckett W Sterner

Database, Volume 2018, 2018, bax100, https://doi.org/10.1093/database/bax100
Published: 04 January 2018 **Article history ▼**

PDF Split View Cite Permissions

Abstract

Growing concerns about the quality of aggregated biodiversity data are lowering trust in large-scale data networks. Aggregators frequently respond to quality concerns by recommending that biologists work with original data providers to correct errors 'at the source.' We show that this strategy falls systematically short of a full diagnosis of the underlying causes of distrust. In particular, trust in an aggregator is not just a feature of the data signal quality provided by the sources to the aggregator, but also a consequence of the social design of the aggregation process and the resulting power balance between individual data contributors and aggregators. The latter have created an accountability gap by downplaying the authorship and significance of the taxonomic hierarchies—frequently called 'backbones'—they generate, and which are in effect novel classification theories that operate at the core of data-structuring process. The Darwin Core standard for sharing occurrence records plays an under-appreciated role in maintaining the accountability gap, because this standard lacks the syntactic structure needed to preserve the taxonomic coherence of data packages submitted for aggregation, potentially leading to inferences that no individual source would support. Since high-quality data packages can mirror competing and conflicting classifications, i.e. unsettled systematic research, this plurality must be accommodated in the design of biodiversity data integration. Looking forward, a key directive is to develop new technical pathways and social incentives for experts to contribute directly to the validation of taxonomically coherent data packages as part of a greater, trustworthy aggregation process.

Issue Section: Perspective/Opinion

- https://doi.org/10.1093/database/bax100
- Critique of monolithic aggregation designs.
- Here we reformulate the issue in relation to the **direction of data flow** in hierarchical biodiversity data aggregation networks.
- Then we outline a complementary solution that is **not centralized,** yet which can potentially reach **global coverage.**

Rethinking centralized biodiversity data aggregation (2)

OXFORD ACADEMIC

DATABASE
The Journal of Biological Databases and Curation

Volumes ▼ Submit ▼ Alerts About ▼ All Database

Volume 2018
2018

Article Contents

Abstract
Introduction
No simple diagnosis for data quality deficiencies
Importance of trust for sustained use
Matching accountability to responsibility
Generation of novel systematic syntheses
Role of the DwC standard
Aggregation and authorship
Consequences of disenfranchizing taxonomic experts
Backbone-based data signal distortion: an example
An objection
Conclusion
Acknowledgements
Funding

To increase trust, change the social design behind aggregated biodiversity data

Nico M Franz ✉, Beckett W Sterner

Database, Volume 2018, 2018, bax100, https://doi.org/10.1093/database/bax100
Published: 04 January 2018 **Article history ▼**

PDF Split View Cite Permissions

Abstract

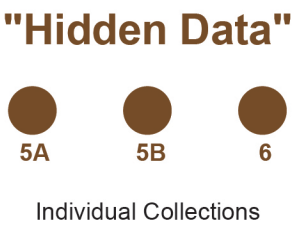
Growing concerns about the quality of aggregated biodiversity data are lowering trust in large-scale data networks. Aggregators frequently respond to quality concerns by recommending that biologists work with original data providers to correct errors 'at the source.' We show that this strategy falls systematically short of a full diagnosis of the underlying causes of distrust. In particular, trust in an aggregator is not just a feature of the data signal quality provided by the sources to the aggregator, but also a consequence of the social design of the aggregation process and the resulting power balance between individual data contributors and aggregators. The latter have created an accountability gap by downplaying the authorship and significance of the taxonomic hierarchies—frequently called 'backbones'—they generate, and which are in effect novel classification theories that operate at the core of data-structuring process. The Darwin Core standard for sharing occurrence records plays an under-appreciated role in maintaining the accountability gap, because this standard lacks the syntactic structure needed to preserve the taxonomic coherence of data packages submitted for aggregation, potentially leading to inferences that no individual source would support. Since high-quality data packages can mirror competing and conflicting classifications, i.e. unsettled systematic research, this plurality must be accommodated in the design of biodiversity data integration. Looking forward, a key directive is to develop new technical pathways and social incentives for experts to contribute directly to the validation of taxonomically coherent data packages as part of a greater, trustworthy aggregation process.

Issue Section: Perspective/Opinion

- https://doi.org/10.1093/database/bax100
- Critique of monolithic aggregation designs.
- Here we reformulate the issue in relation to the **direction of data flow** in hierarchical biodiversity data aggregation networks.
- Then we outline a complementary solution that is **not centralized,** yet which can potentially reach **global coverage.**

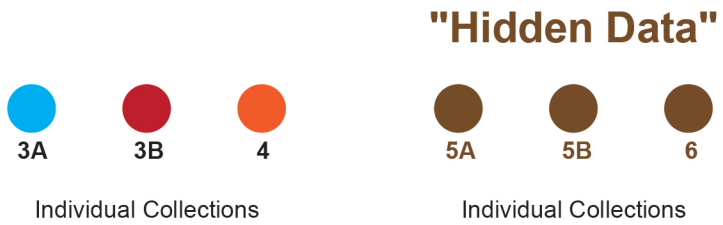
Centralized biodiversity data aggregation

- Many collections remain offline.



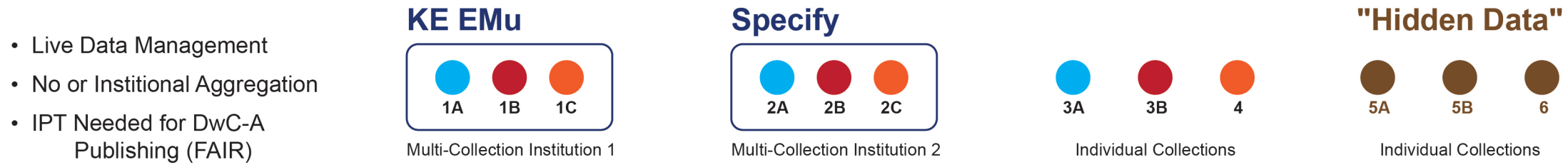
Centralized biodiversity data aggregation

- Other, individual collections, may have a web presence.



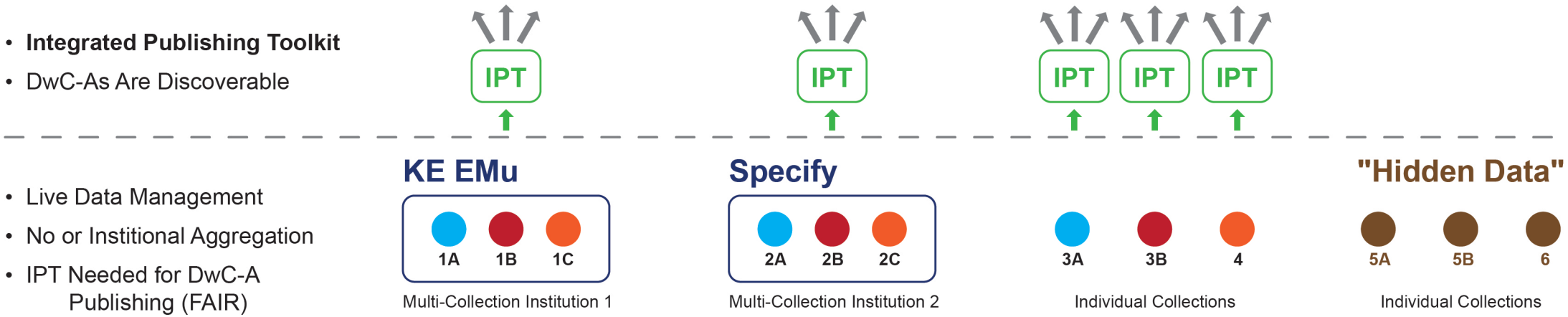
Centralized biodiversity data aggregation

- Multi-collections software applications can bring institutional datasets on-line; yet this alone may not suffice to **publish DwC-Archive data** according to [FAIR standards.]



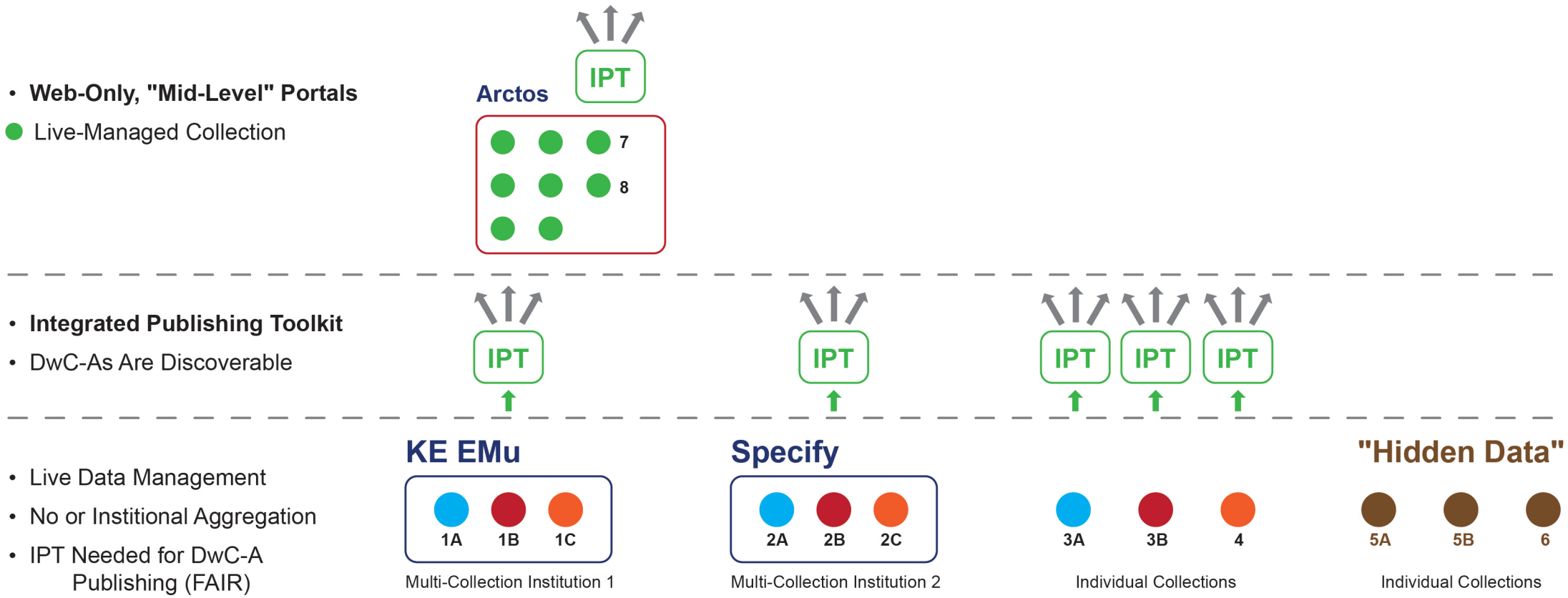
Centralized biodiversity data aggregation

- The Integrated Publishing Toolkit (IPT) allows individual or multi-collection datasets to **become discoverable** as DwC-Archive packages to higher-level aggregators.



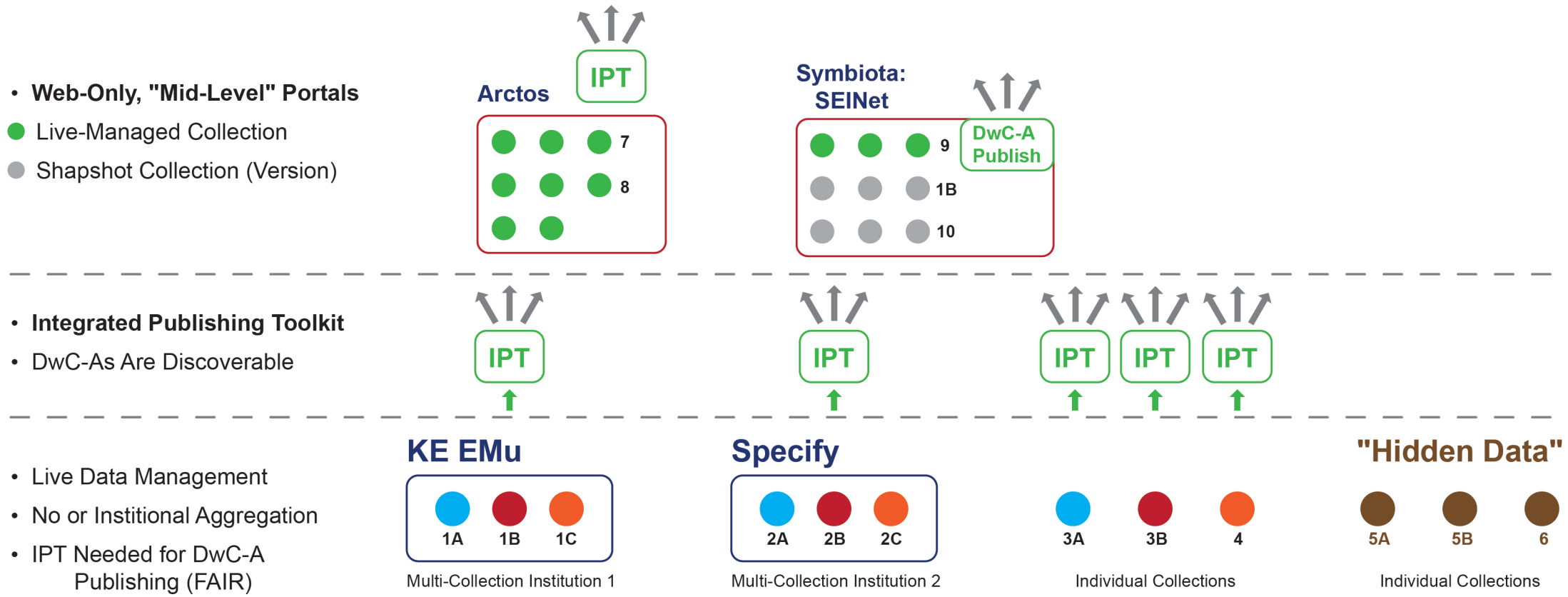
Centralized biodiversity data aggregation

- There are also **web-only, mid-level portal applications** that support live collection management and can publish "up" through the (IPT).



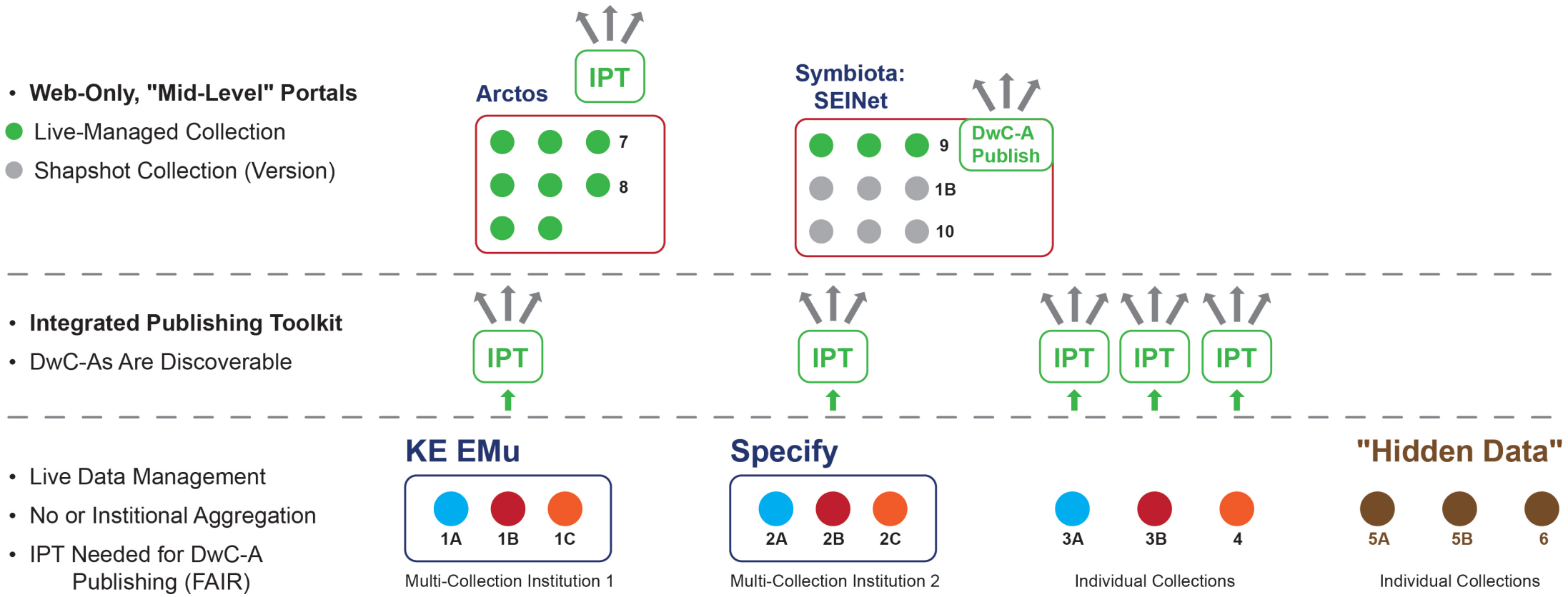
Centralized biodiversity data aggregation

- Symbiota portals also support **"snapshot collections"** – i.e., periodical, manually triggered batch re-/uploads of static versions that are live-managed elsewhere.

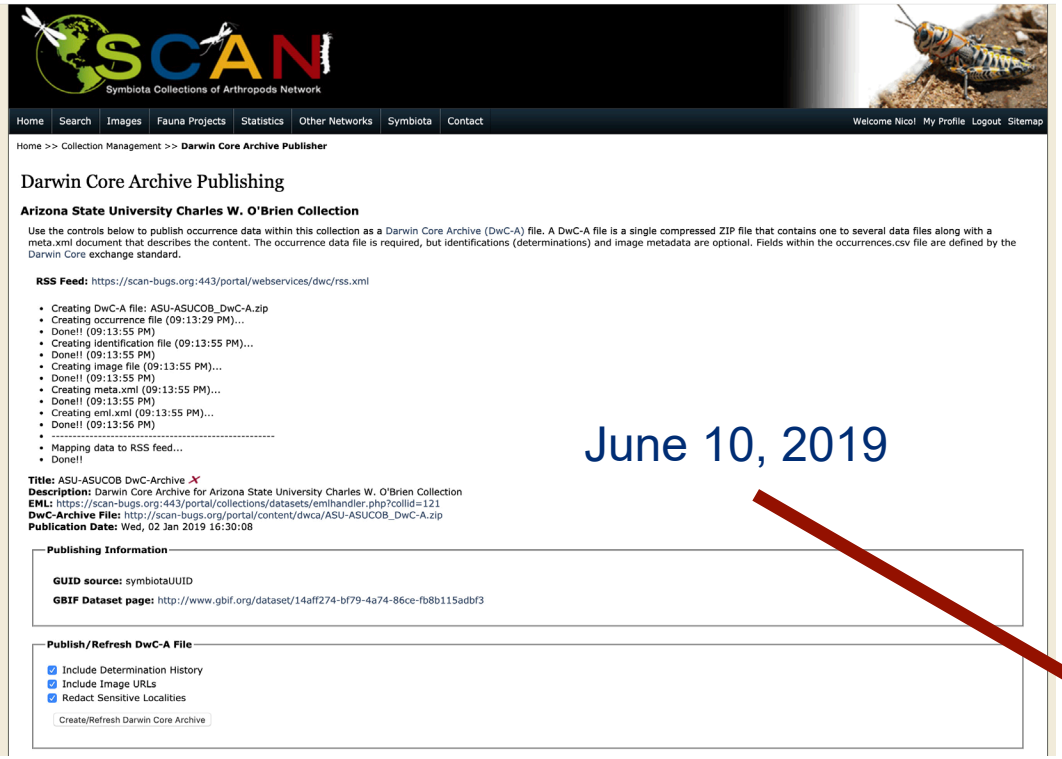


Centralized biodiversity data aggregation

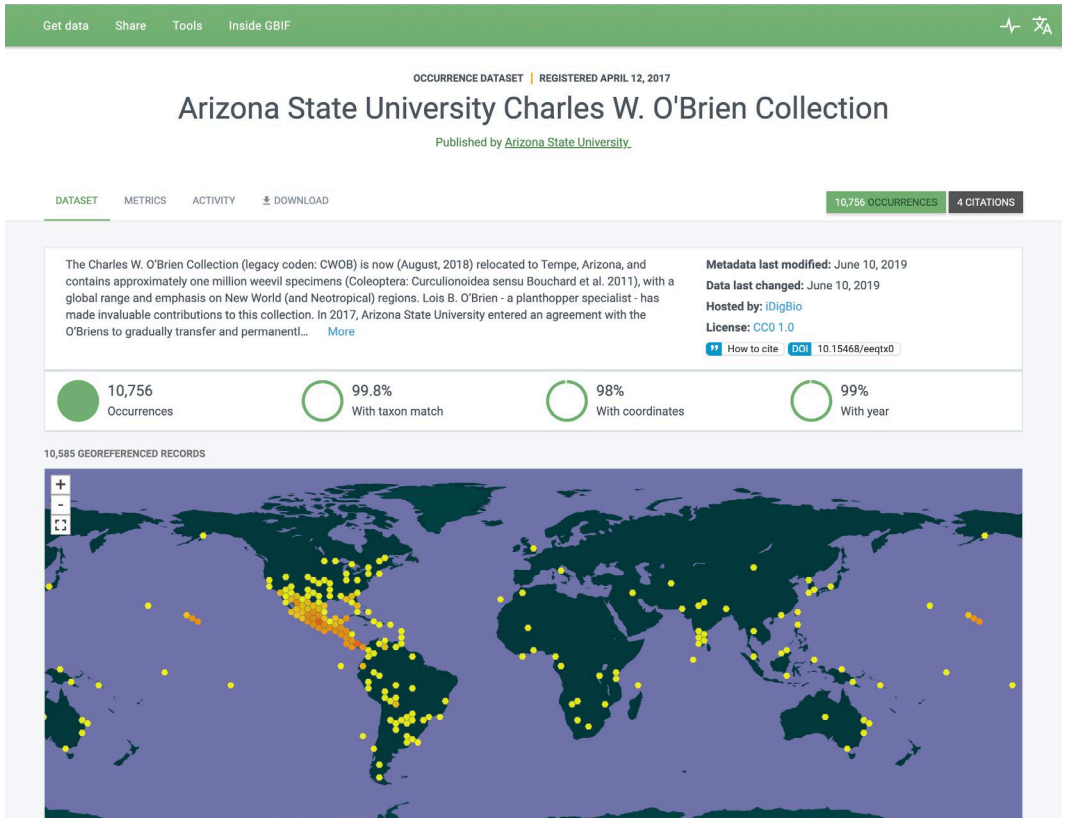
- Symbiota portals also have a custom, fully built-in, IPT-analogous **"Darwin Core Archive Publishing"** module.



Example of Symbiota's DwC-A Publishing module (SCAN: ASUCOB)

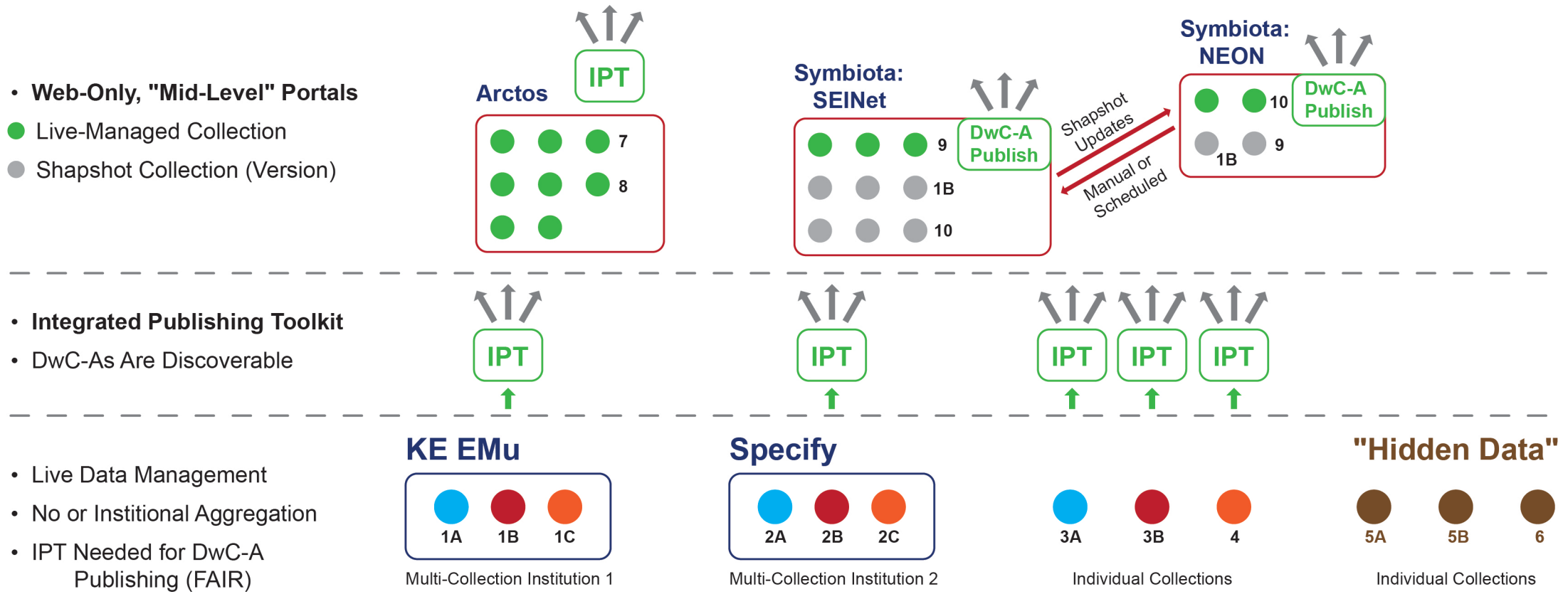


June 10, 2019

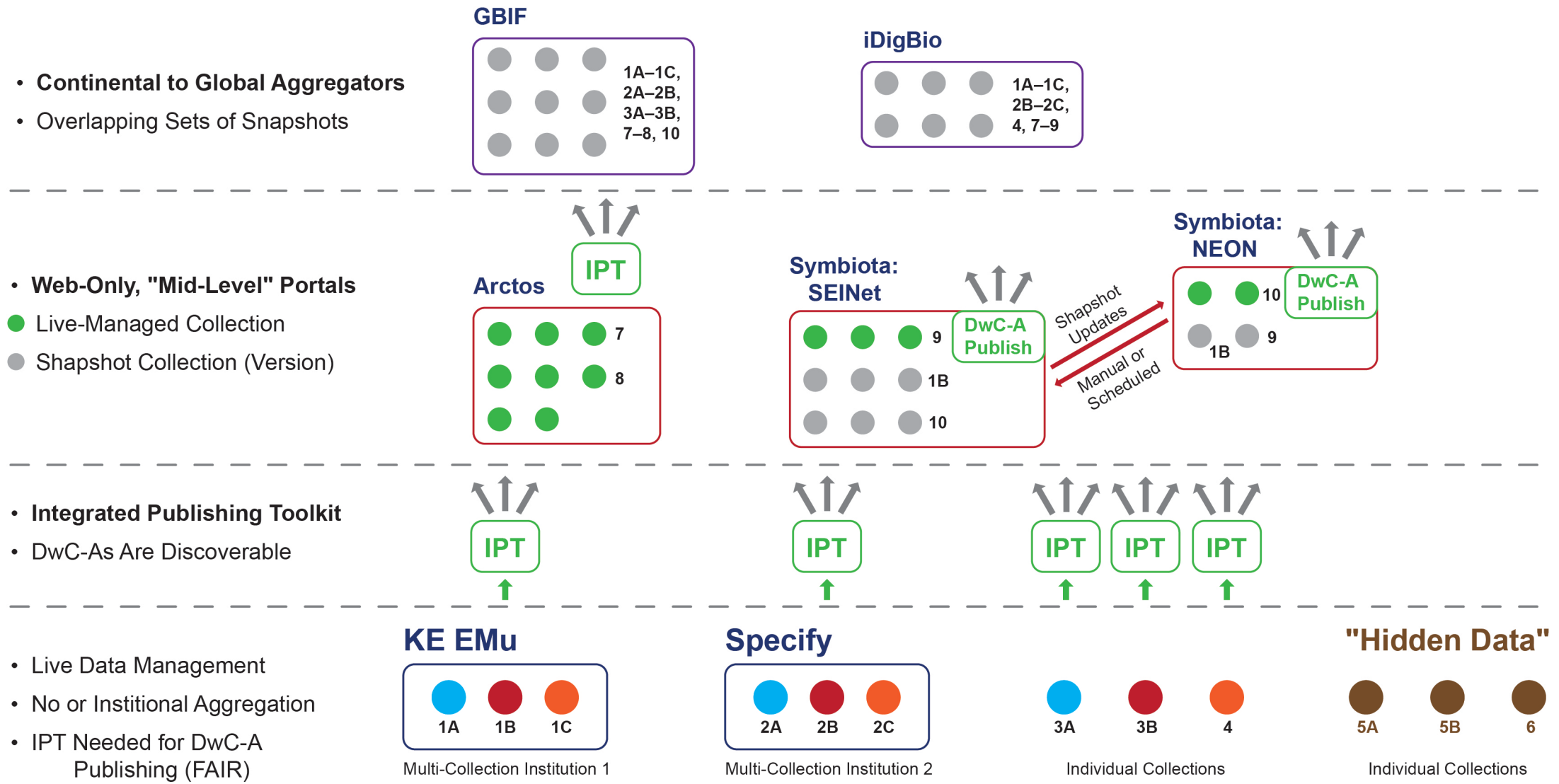


Centralized biodiversity data aggregation

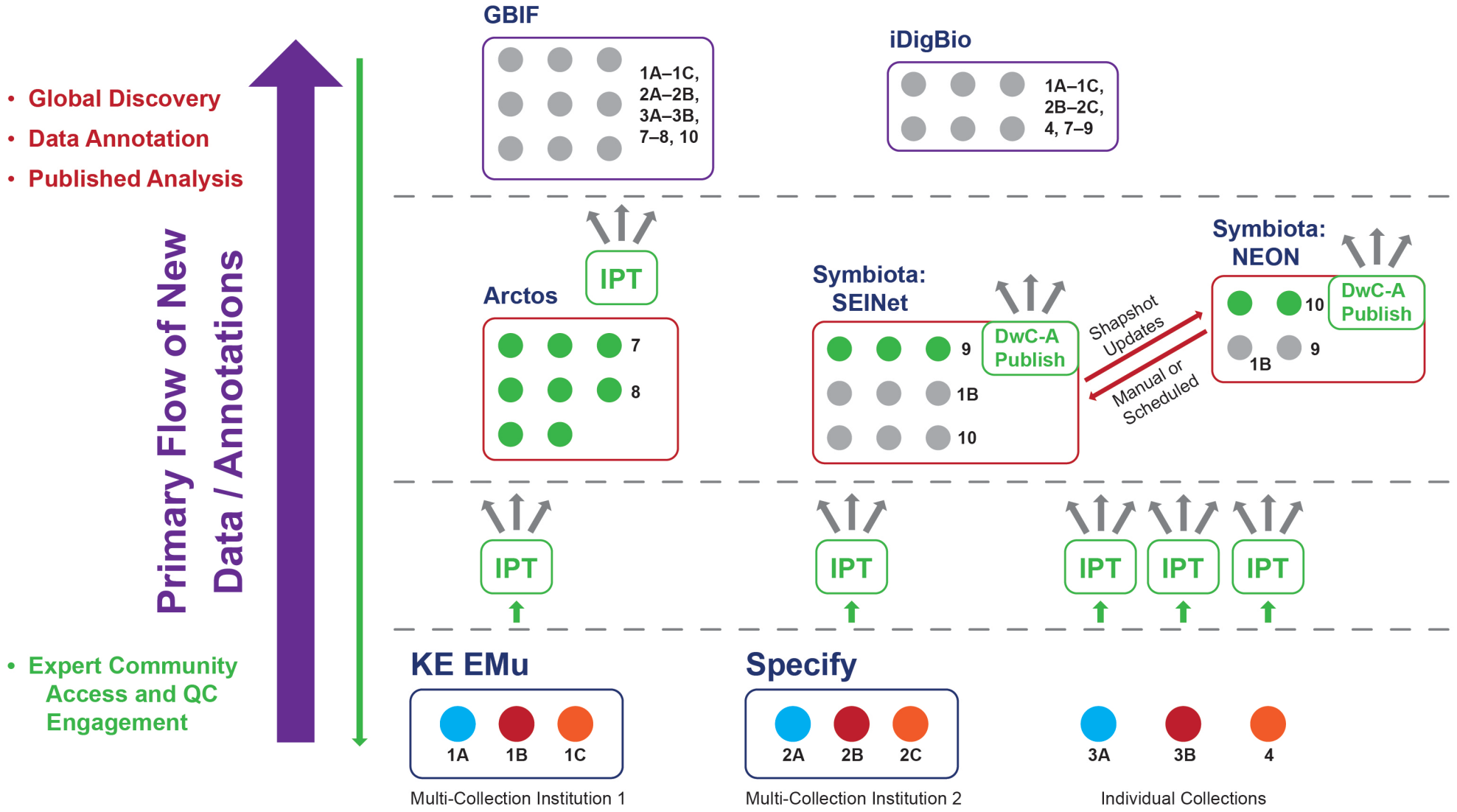
- Multiple, community-themed portals – each with unique live/snapshot collection profiles – can periodically receive **reciprocal snapshot updates.**



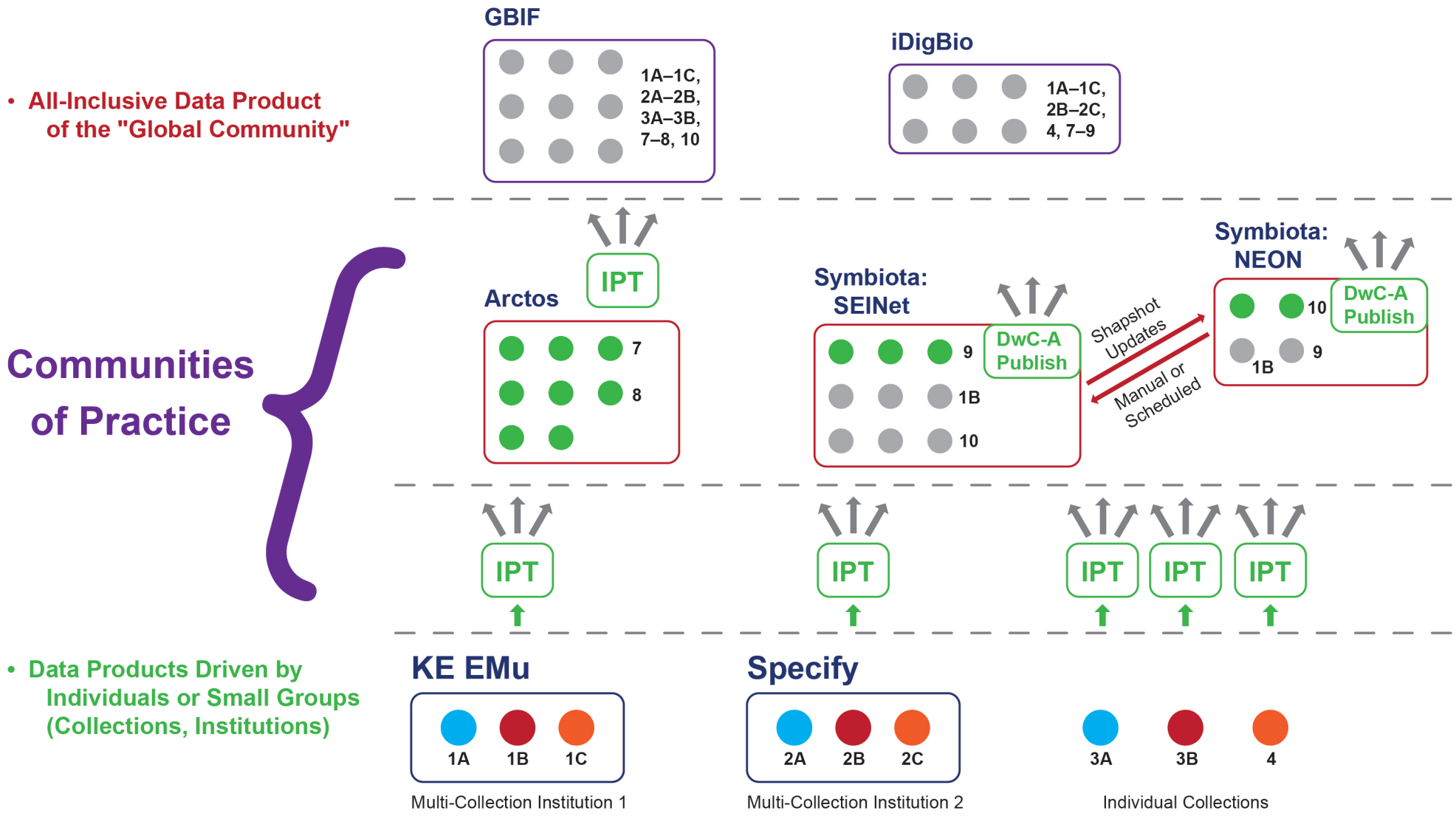
Highest-level aggregators typically only support collection **snapshots!**



This hierarchy sustains an **imbalance in directional data flow:**
Annotations on global datasets are hard to pull downwards.



Moreover, by the time we reach the top, most **experts/enthusiasts no longer feel at home** (cf. Wenger 2000).

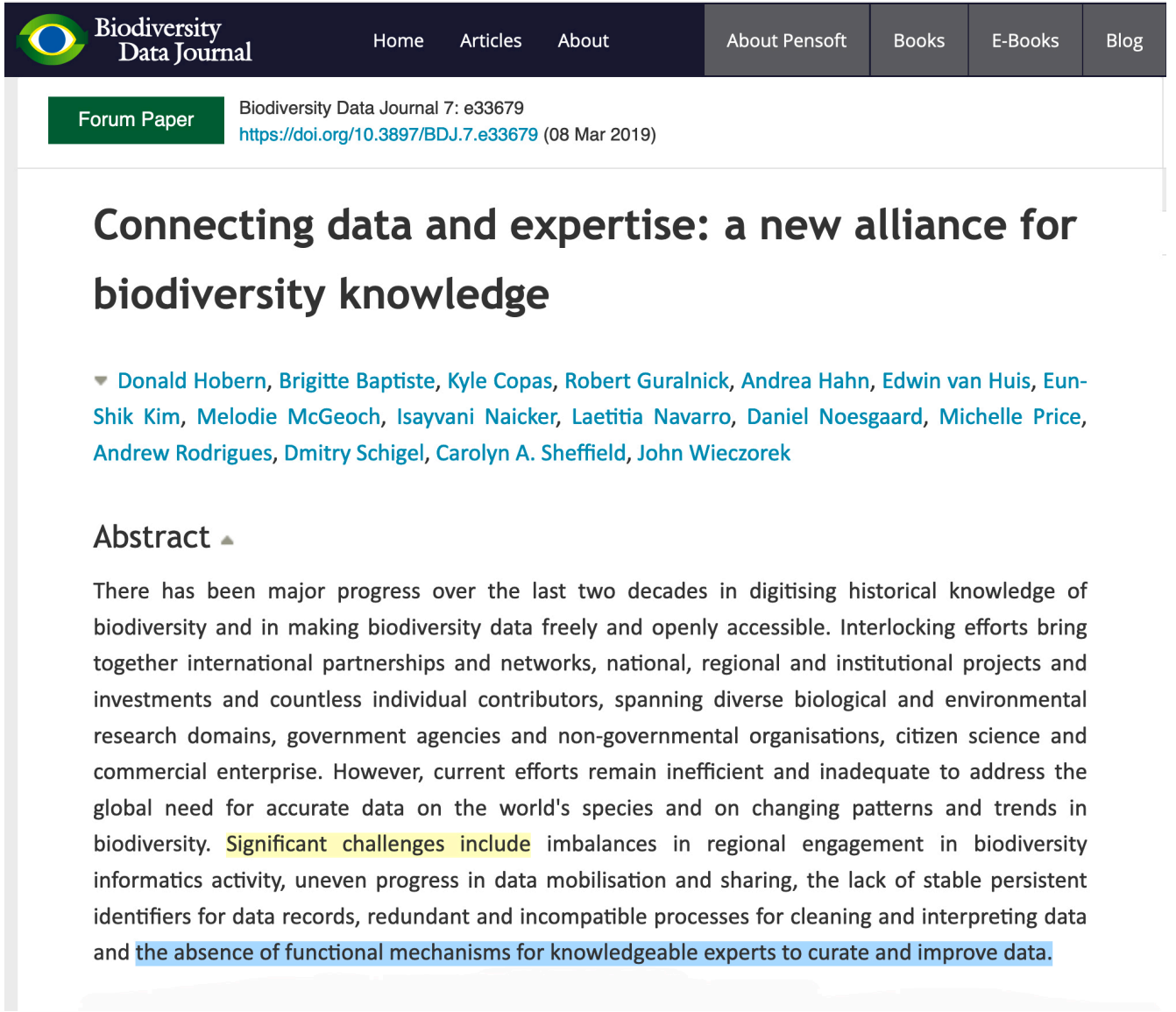


Wenger 2000: Communities of practice and social learning systems

Table 2. Boundary Dimensions

	Coordination	Transparency	Negotiability
Engagement	What opportunities exist for joint activities, problem-solving, and discussions to both surface and resolve differences through action?	Do people provide explanations, coaching, and demonstrations in the context of joint activities to open windows on to each other's practices?	Are joint activities structured in such a way that multiple perspectives can meet and participants can come to appreciate each other's competences?
Imagination	Do people have enough understanding of their respective perspectives to present issues effectively and anticipate misunderstandings?	What stories, documents, and models are available to build a picture of another practice? What experience will allow people to walk in the other's shoes? Do they listen deeply enough?	Can both sides see themselves as members of an overarching community in which they have common interests and needs?
Alignment	Are instructions, goals, and methods interpretable into action across boundaries?	Are intentions, commitments, norms, and traditions made clear enough to reveal common ground and differences in perspectives and expectations?	Who has a say in negotiating contracts and devising compromises?

Community Access ⇔ Engagement ⇔ Quality ⇔ Trust ⇔ Use & Impact



Biodiversity Data Journal

Home Articles About About Pensoft Books E-Books Blog

Forum Paper

Biodiversity Data Journal 7: e33679
https://doi.org/10.3897/BDJ.7.e33679 (08 Mar 2019)

Connecting data and expertise: a new alliance for biodiversity knowledge

Donald Hobern, Brigitte Baptiste, Kyle Copas, Robert Guralnick, Andrea Hahn, Edwin van Huis, Eun-Shik Kim, Melodie McGeoch, Isayvani Naicker, Laetitia Navarro, Daniel Noesgaard, Michelle Price, Andrew Rodrigues, Dmitry Schigel, Carolyn A. Sheffield, John Wieczorek

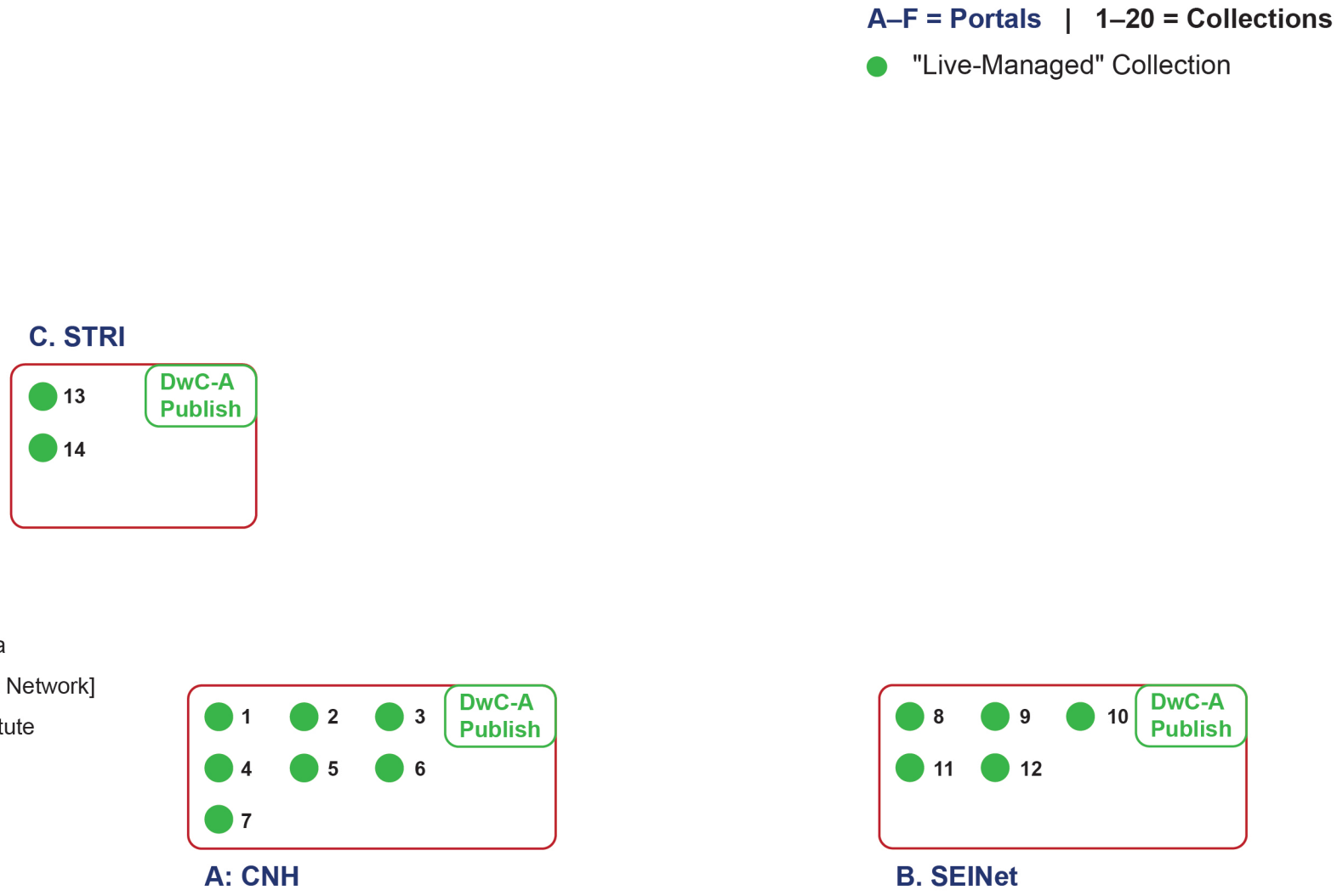
Abstract

There has been major progress over the last two decades in digitising historical knowledge of biodiversity and in making biodiversity data freely and openly accessible. Interlocking efforts bring together international partnerships and networks, national, regional and institutional projects and investments and countless individual contributors, spanning diverse biological and environmental research domains, government agencies and non-governmental organisations, citizen science and commercial enterprise. However, current efforts remain inefficient and inadequate to address the global need for accurate data on the world's species and on changing patterns and trends in biodiversity. Significant challenges include imbalances in regional engagement in biodiversity informatics activity, uneven progress in data mobilisation and sharing, the lack of stable persistent identifiers for data records, redundant and incompatible processes for cleaning and interpreting data and the absence of functional mechanisms for knowledgeable experts to curate and improve data.

Designing for strong data communities

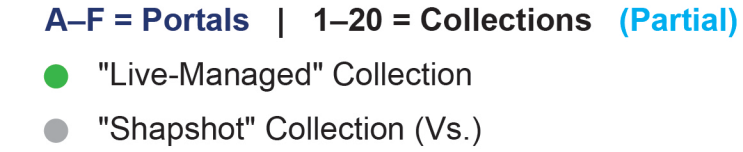
De-centralized, but global

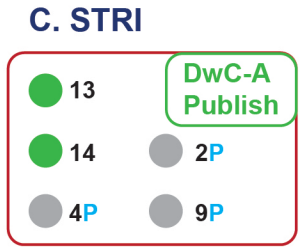
- Independent, themed portal communities maintain **live-managed collections.**



De-centralized, but global

- **Partial,** **relevant *collection* snapshot** **subsets** are represented.

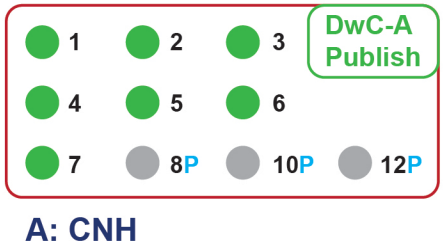


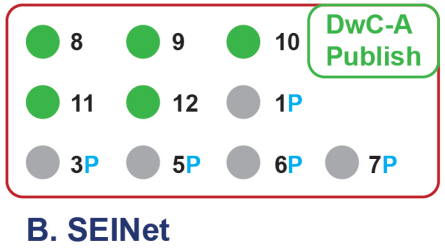


A: Consortium of Northeastern Herbaria

B: [Southwest Environment Information Network]

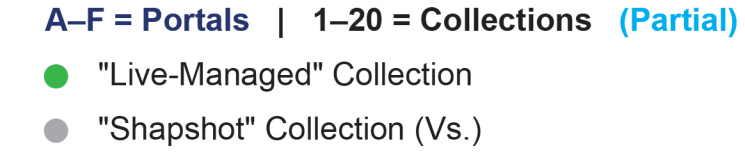
C: Smithsonian Tropical Research Institute

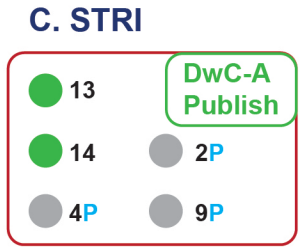


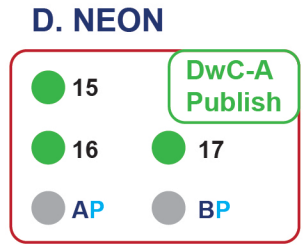


De-centralized, but global

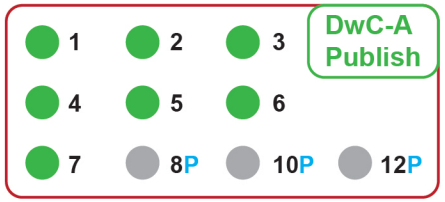
- Even **partial,** **relevant *portal* snapshot subsets** are ingestible, with provenance.

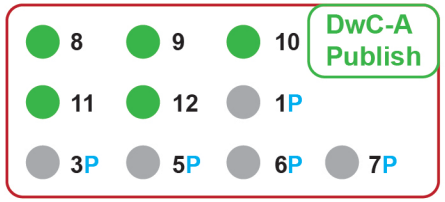






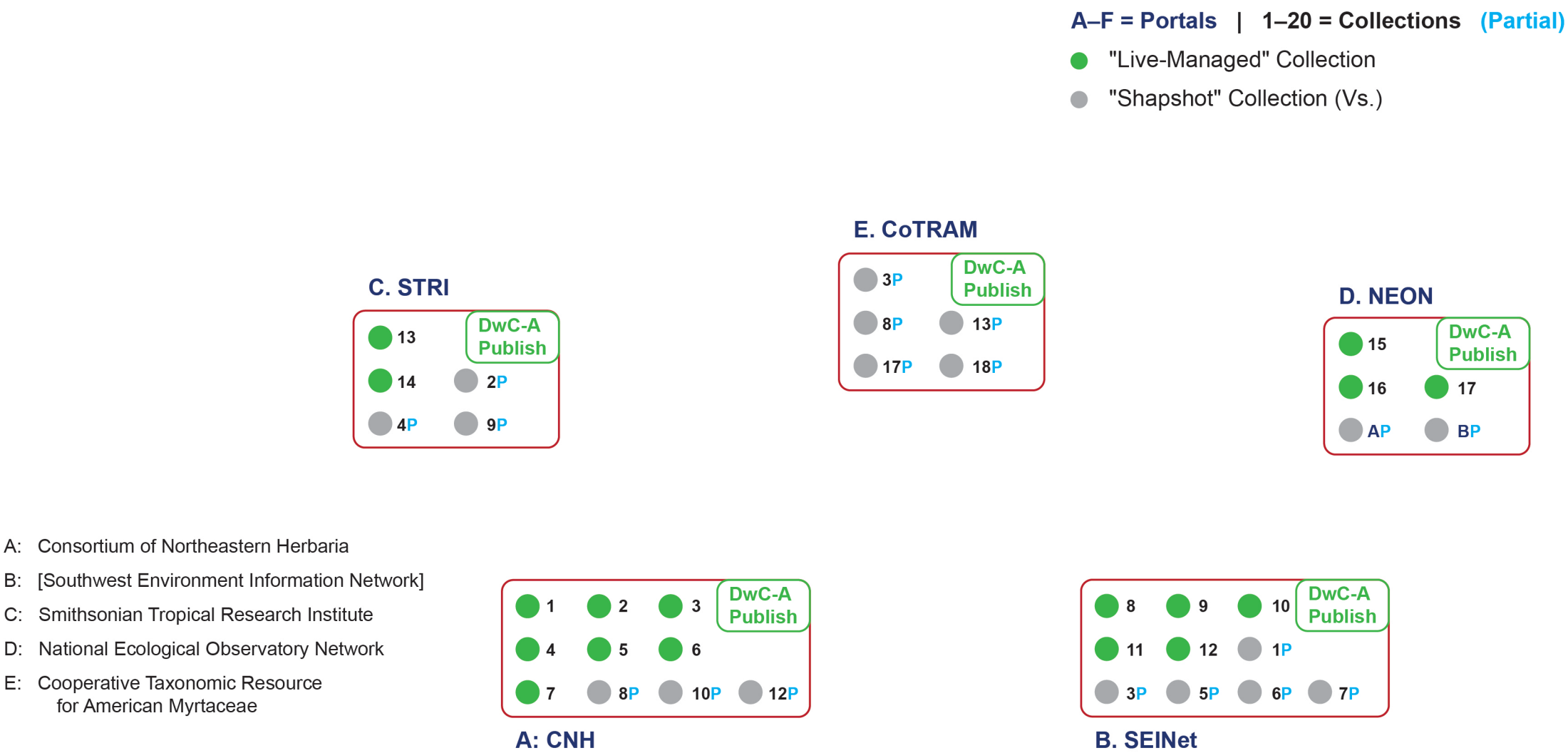
A: Consortium of Northeastern Herbaria
B: [Southwest Environment Information Network]
C: Smithsonian Tropical Research Institute
D: National Ecological Observatory Network





De-centralized, but global

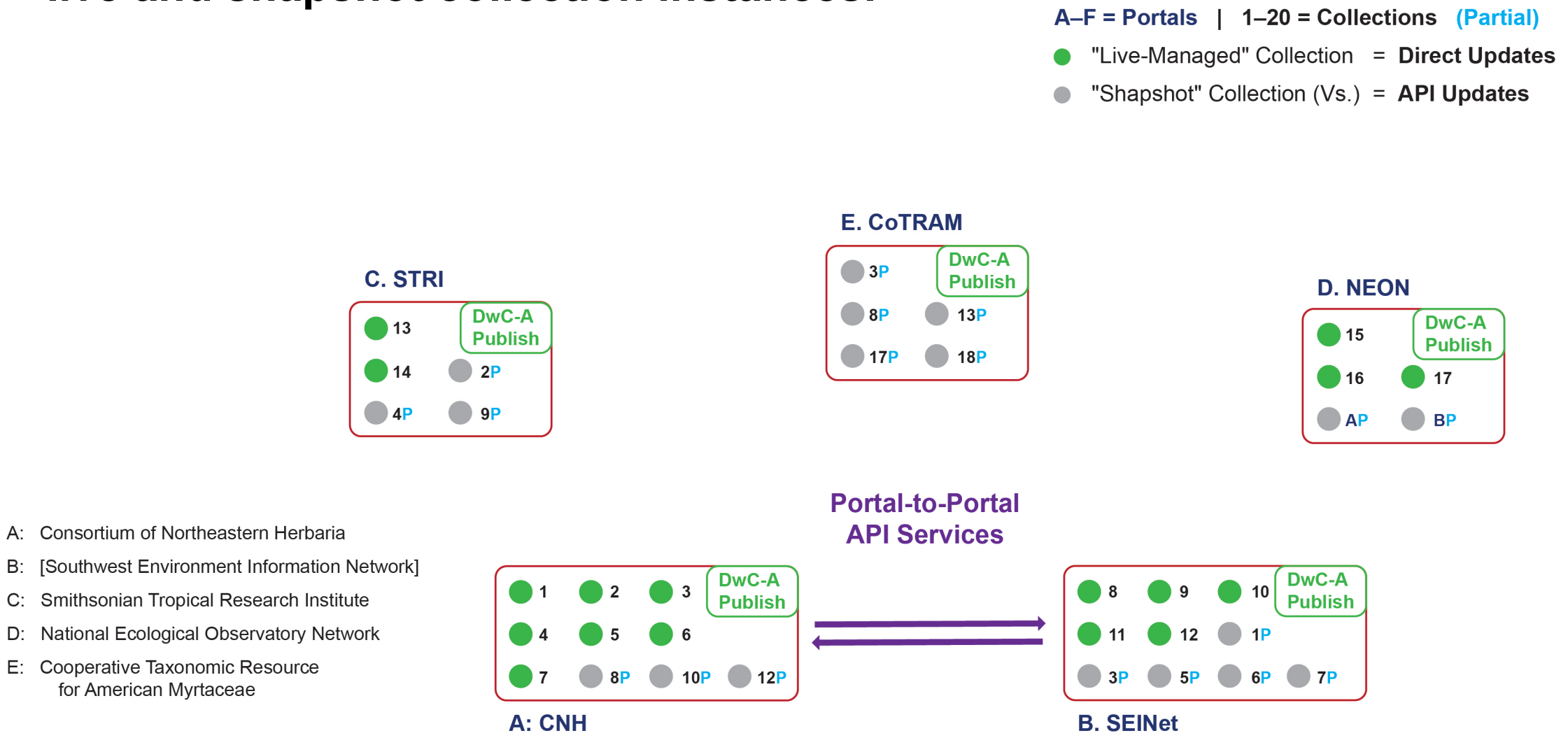
- Some research-themed portals may only include partial collection snapshots.



A: Consortium of Northeastern Herbaria
B: [Southwest Environment Information Network]
C: Smithsonian Tropical Research Institute
D: National Ecological Observatory Network
E: Cooperative Taxonomic Resource for American Myrtaceae

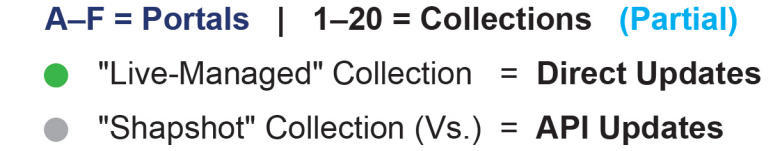
De-centralized, but global

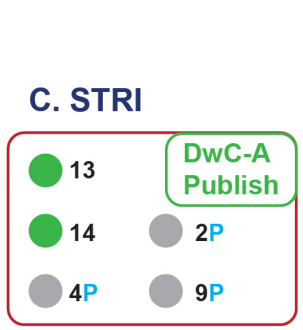
- Highly configurable **portal-to-portal APIs** negotiate the flow of data **between live and snapshot collection instances.**

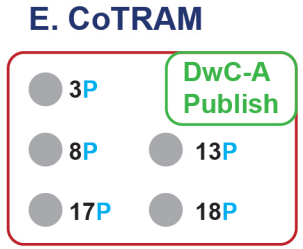


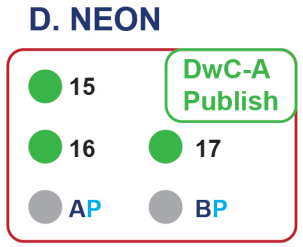
De-centralized, but global

- As API services are optimized, the **distinction between live and snapshot** collection management **increasingly falls away.**

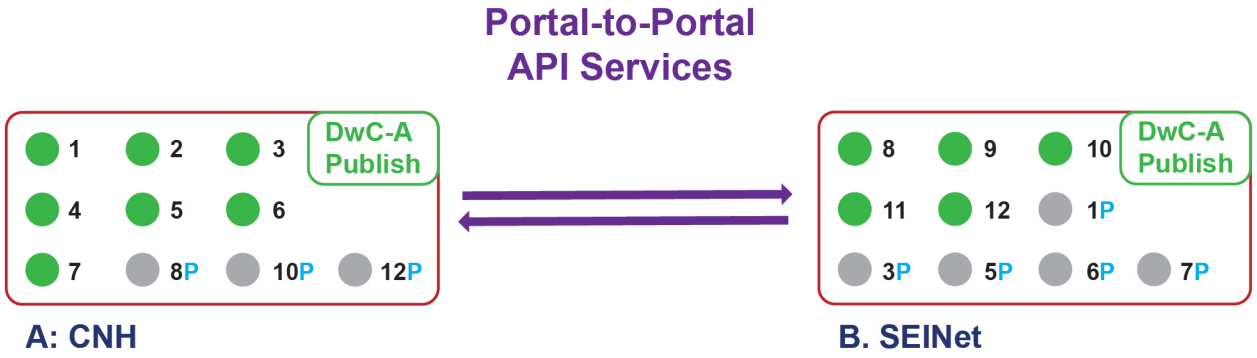






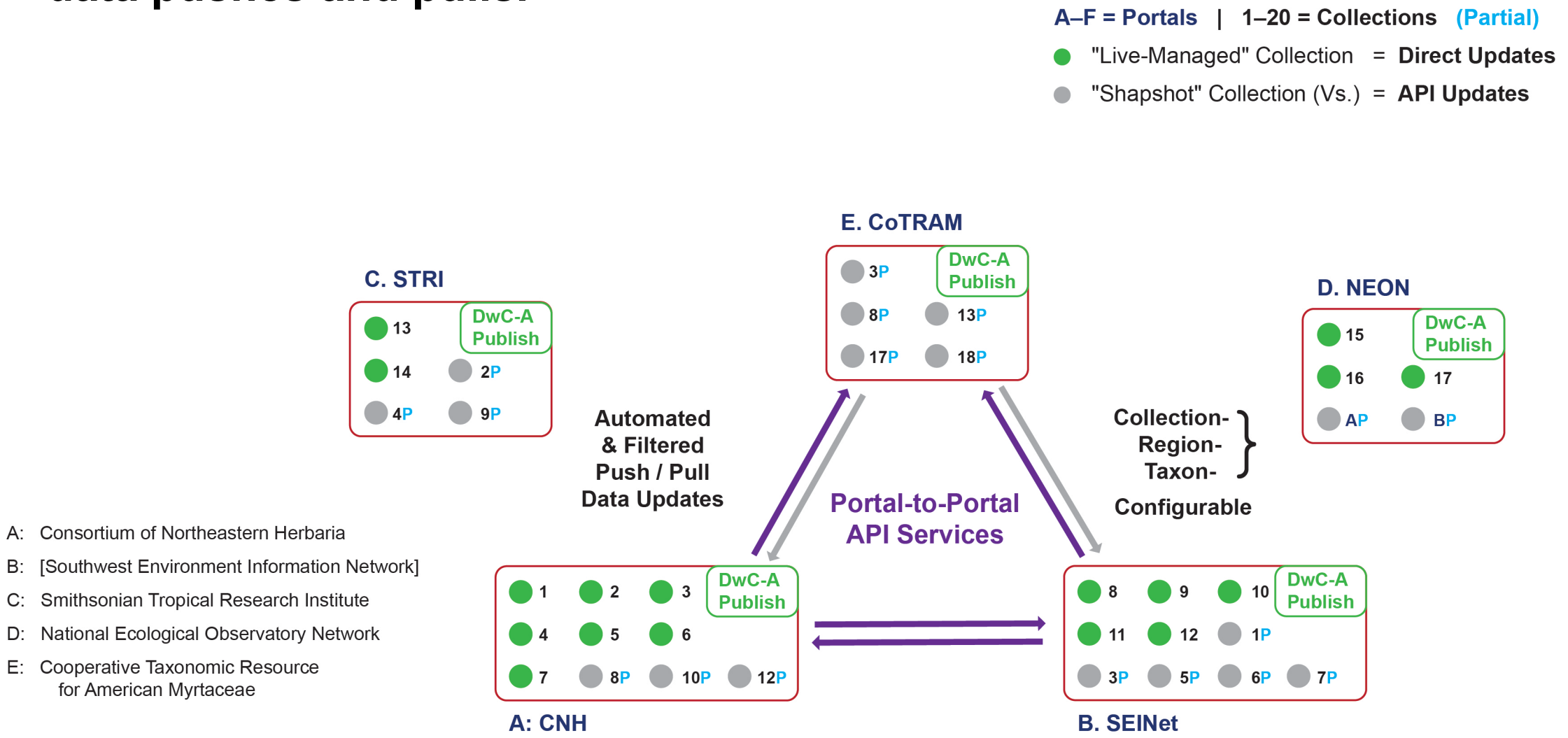


A: Consortium of Northeastern Herbaria
B: [Southwest Environment Information Network]
C: Smithsonian Tropical Research Institute
D: National Ecological Observatory Network
E: Cooperative Taxonomic Resource for American Myrtaceae



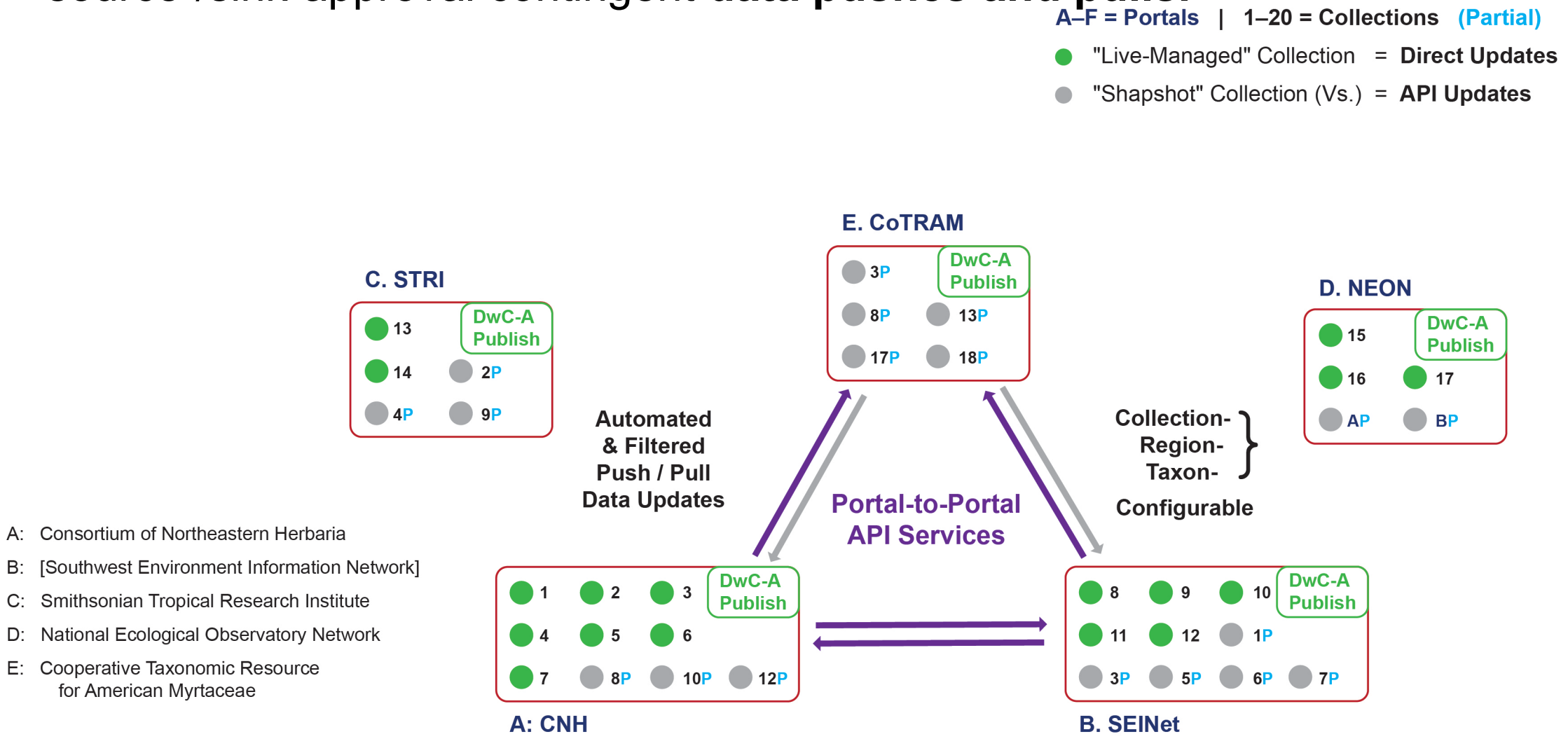
De-centralized, but global

- API service configurations include **filtered,** source-/sink-approval contingent **data pushes and pulls.**



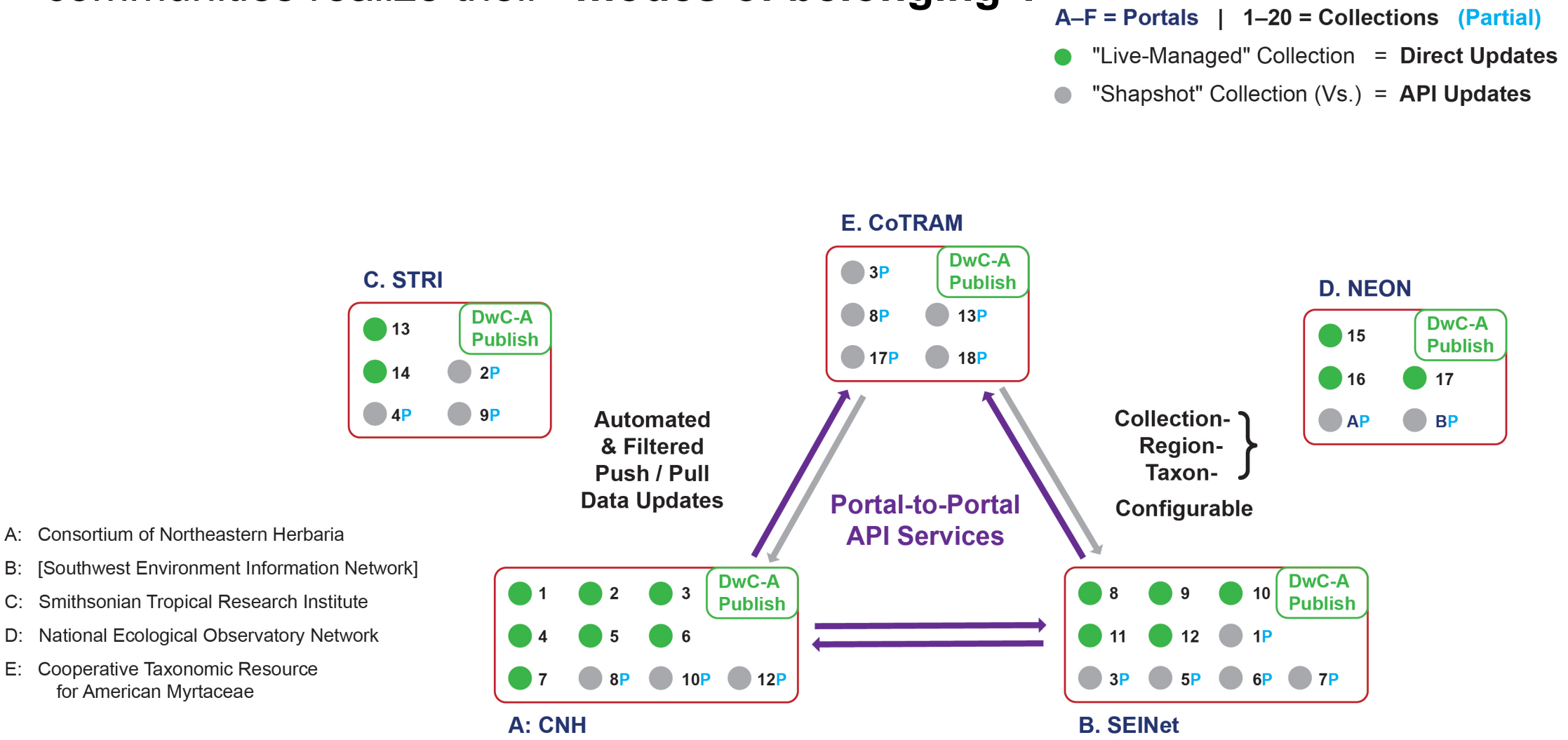
De-centralized, but global

- API service configurations allow **filtered** {collection, taxon, region, etc.)**,** source-/sink-approval contingent **data pushes and pulls.**



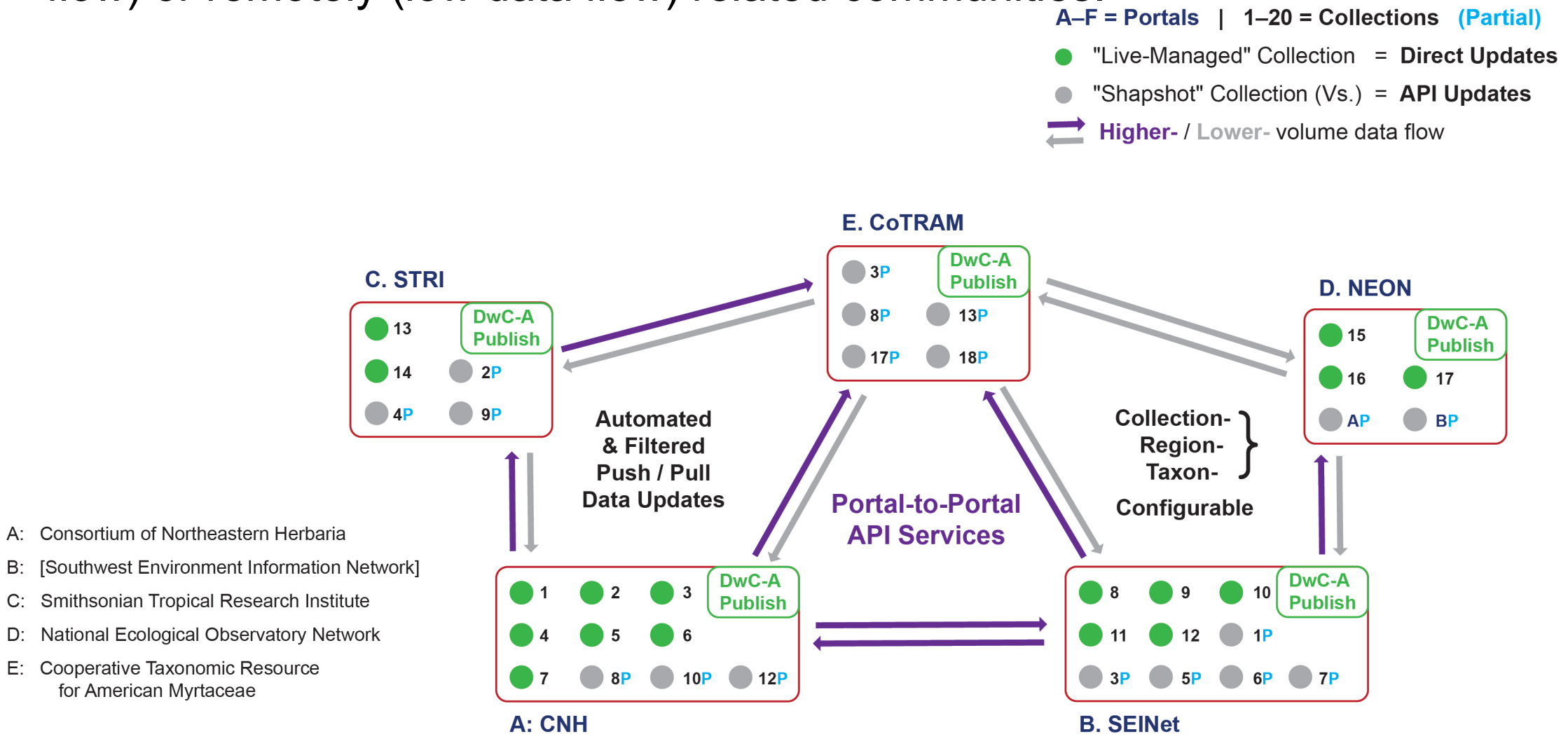
De-centralized, but global

- Portal-to-portal **API configurations** become the "substrate" upon which the communities realize their **"modes of belonging".**



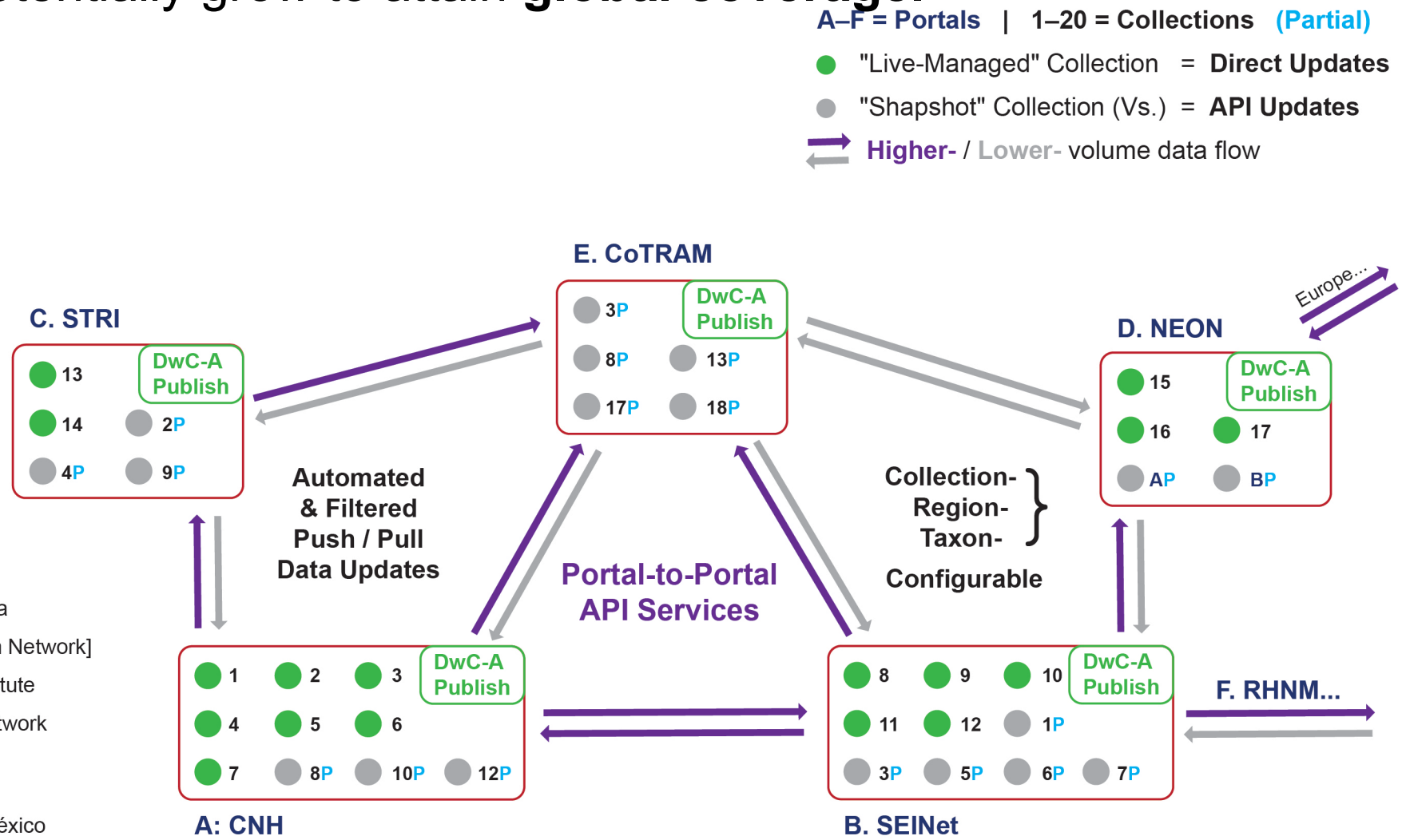
De-centralized, but global

- The de-centralized network is **broadly extensible** between closely (high data flow) or remotely (low data flow) related communities.



De-centralized, but global

- On the basis of a shared API service culture, a de-centralized data portal network can potentially grow to attain **global coverage.**

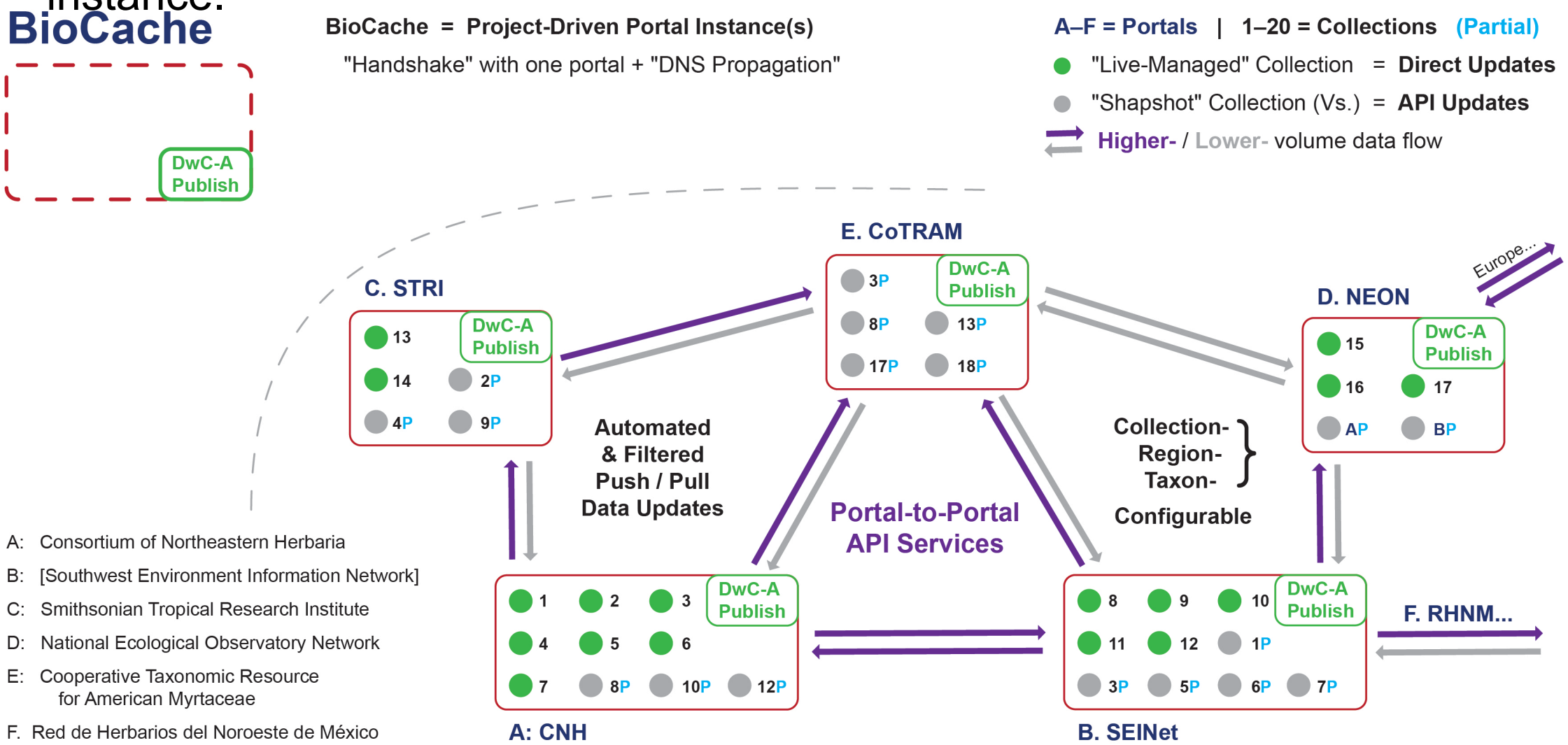


A: Consortium of Northeastern Herbaria
B: [Southwest Environment Information Network]
C: Smithsonian Tropical Research Institute
D: National Ecological Observatory Network
E: Cooperative Taxonomic Resource for American Myrtaceae
F. Red de Herbarios del Noroeste de México

Designing for expert/enthusiast access

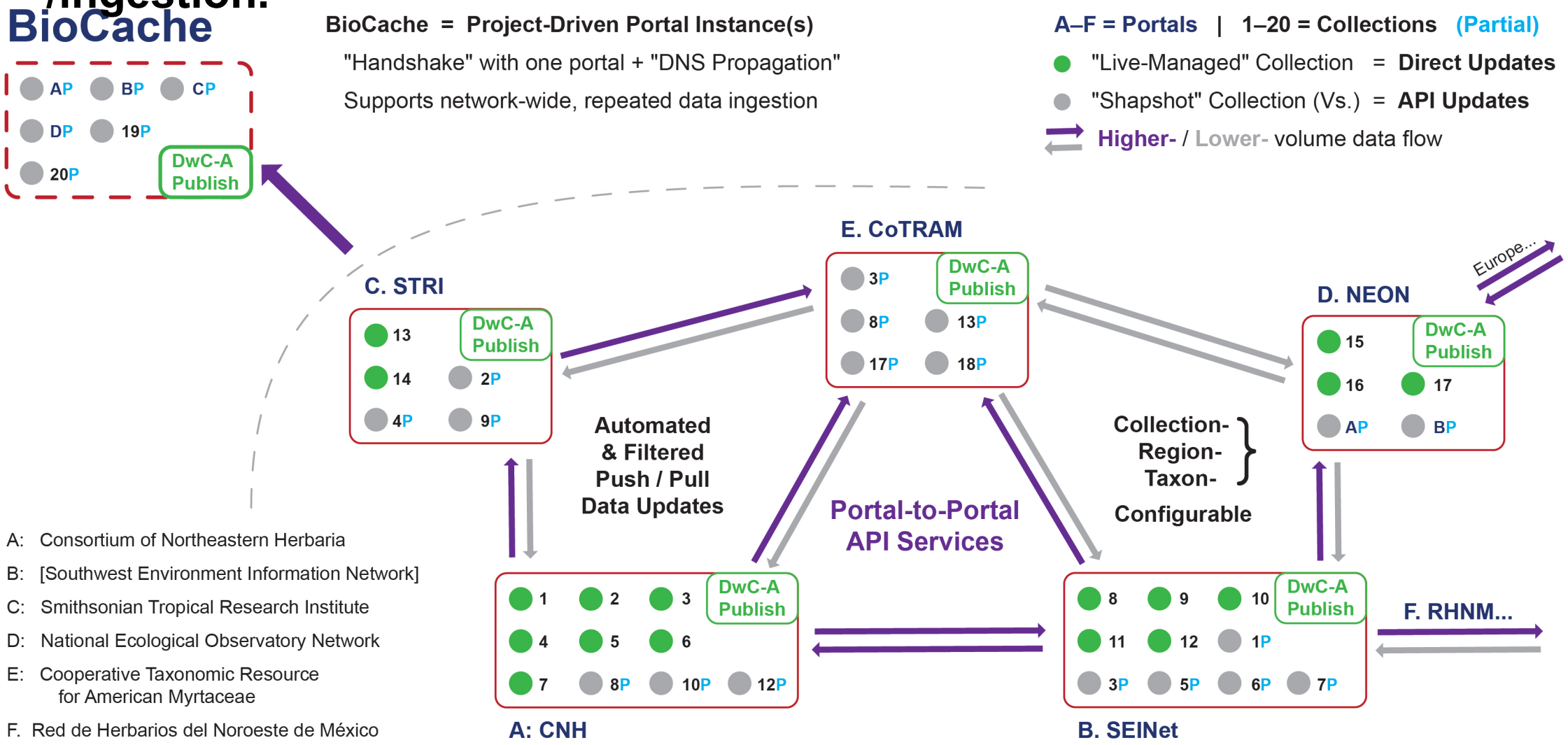
BioCache: Global access through **custom research portal instances**

- Researchers create and **register "via single handshake"** a new portal instance.



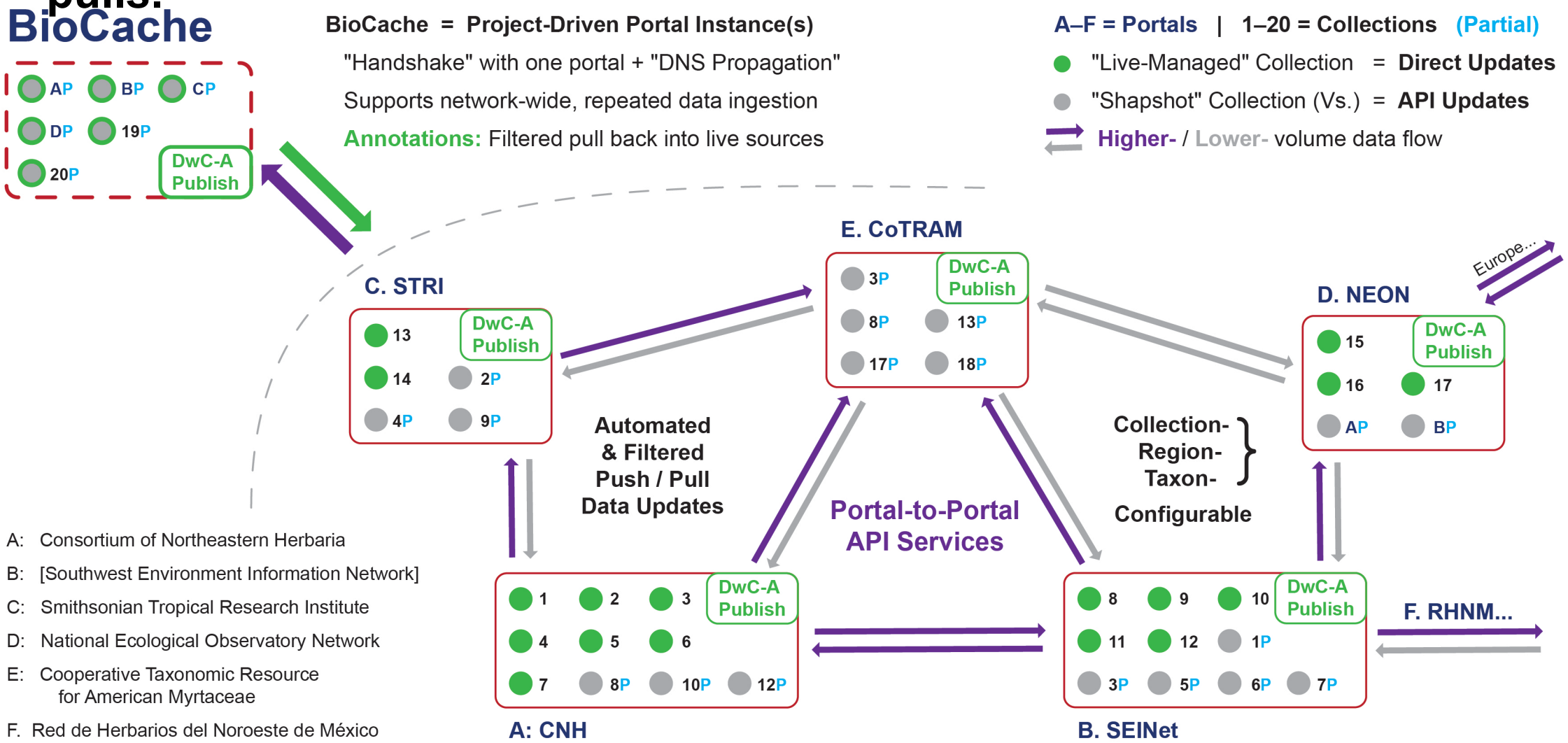
BioCache: Global access through **custom research portal instances**

- Research instances enable **repeatable, global data queries and re-/ingestion.**



BioCache: Global access through **custom research portal instances**

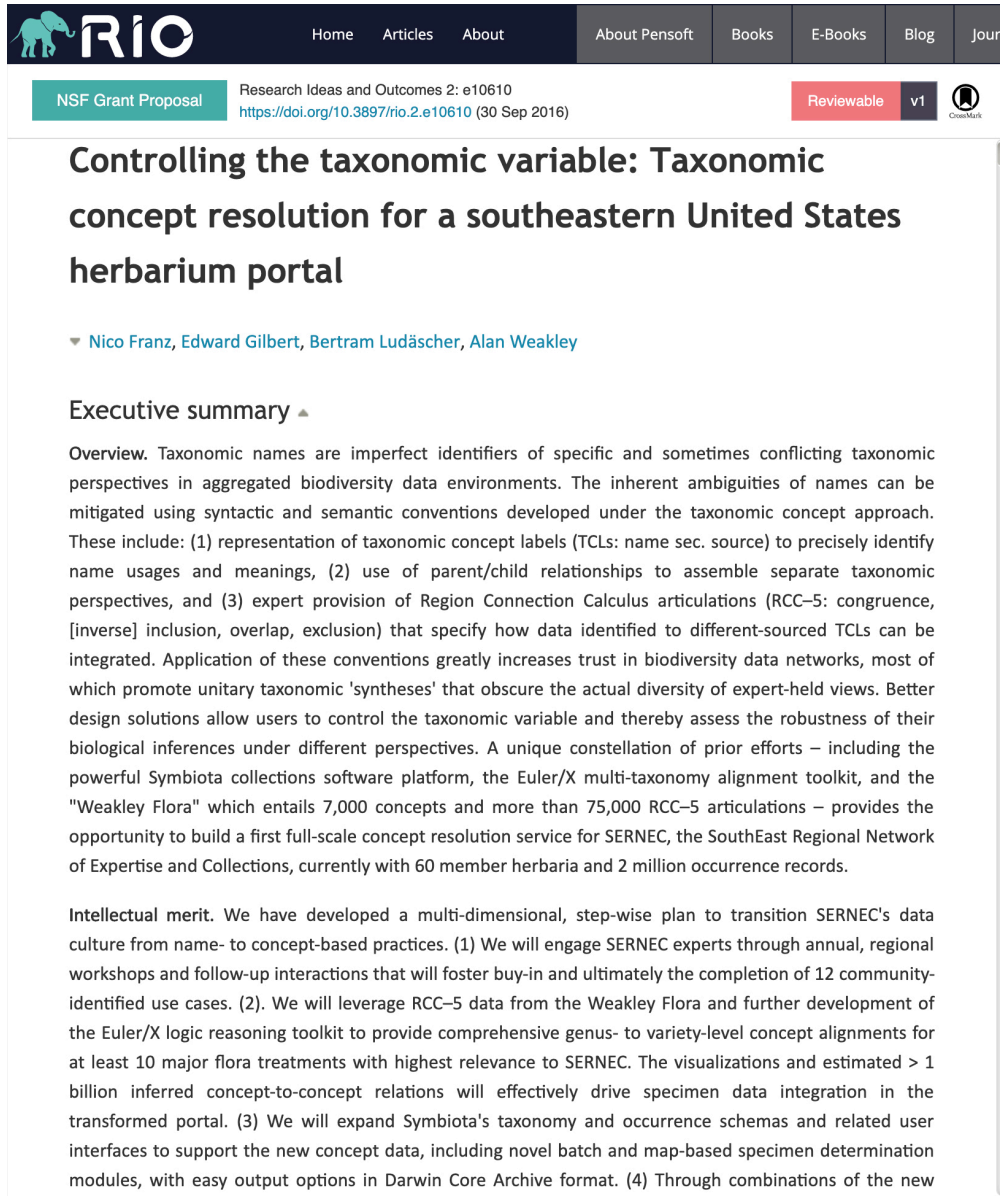
- **Valued-added data can return** to all (live) source collections **via filtered pulls.**



A: Consortium of Northeastern Herbaria
B: [Southwest Environment Information Network]
C: Smithsonian Tropical Research Institute
D: National Ecological Observatory Network
E: Cooperative Taxonomic Resource for American Myrtaceae
F. Red de Herbarios del Noroeste de México

Stay tuned, it's underway

Taxonomic data intelligence for Darwin Core occurrences **(3)**



RIO Home Articles About | About Pensoft Books E-Books Blog Jour

NSF Grant Proposal

Research Ideas and Outcomes 2: e10610
https://doi.org/10.3897/rio.2.e10610 (30 Sep 2016)

Reviewable v1

Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern United States herbarium portal

Nico Franz, Edward Gilbert, Bertram Ludäscher, Alan Weakley

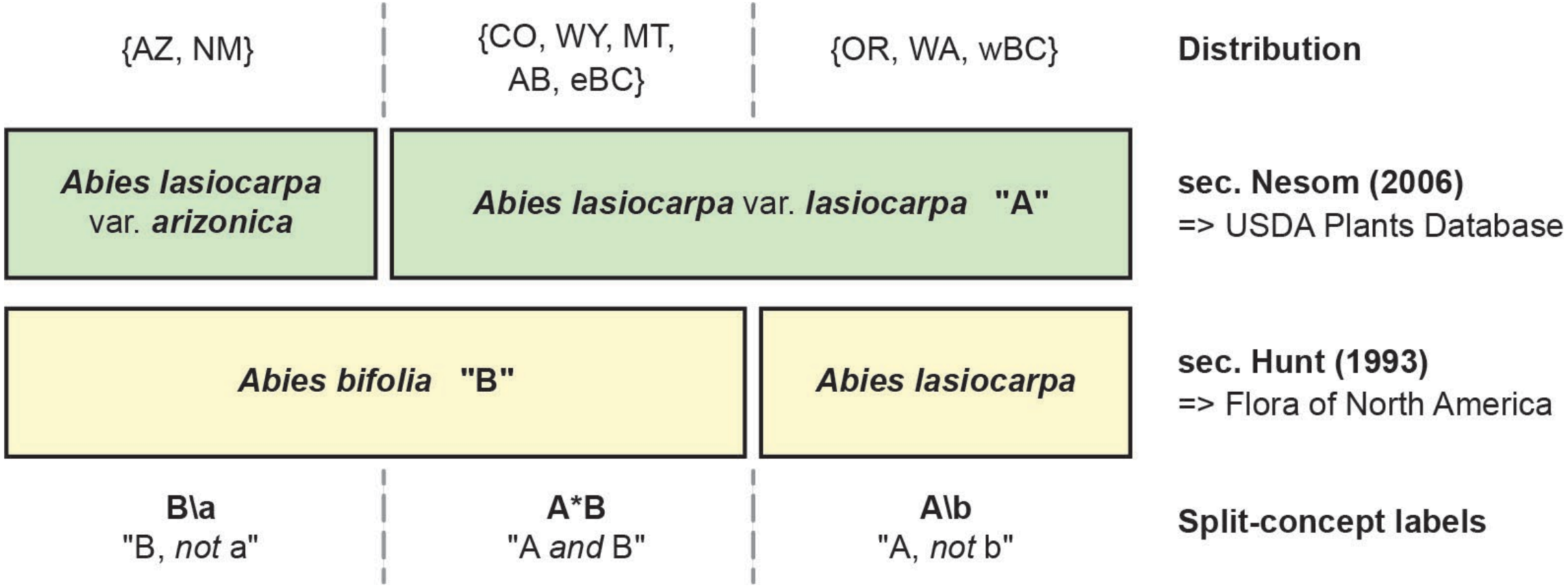
Executive summary

Overview. Taxonomic names are imperfect identifiers of specific and sometimes conflicting taxonomic perspectives in aggregated biodiversity data environments. The inherent ambiguities of names can be mitigated using syntactic and semantic conventions developed under the taxonomic concept approach. These include: (1) representation of taxonomic concept labels (TCLs: name sec. source) to precisely identify name usages and meanings, (2) use of parent/child relationships to assemble separate taxonomic perspectives, and (3) expert provision of Region Connection Calculus articulations (RCC–5: congruence, [inverse] inclusion, overlap, exclusion) that specify how data identified to different-sourced TCLs can be integrated. Application of these conventions greatly increases trust in biodiversity data networks, most of which promote unitary taxonomic 'syntheses' that obscure the actual diversity of expert-held views. Better design solutions allow users to control the taxonomic variable and thereby assess the robustness of their biological inferences under different perspectives. A unique constellation of prior efforts – including the powerful Symbiota collections software platform, the Euler/X multi-taxonomy alignment toolkit, and the "Weakley Flora" which entails 7,000 concepts and more than 75,000 RCC–5 articulations – provides the opportunity to build a first full-scale concept resolution service for SERNEC, the SouthEast Regional Network of Expertise and Collections, currently with 60 member herbaria and 2 million occurrence records.

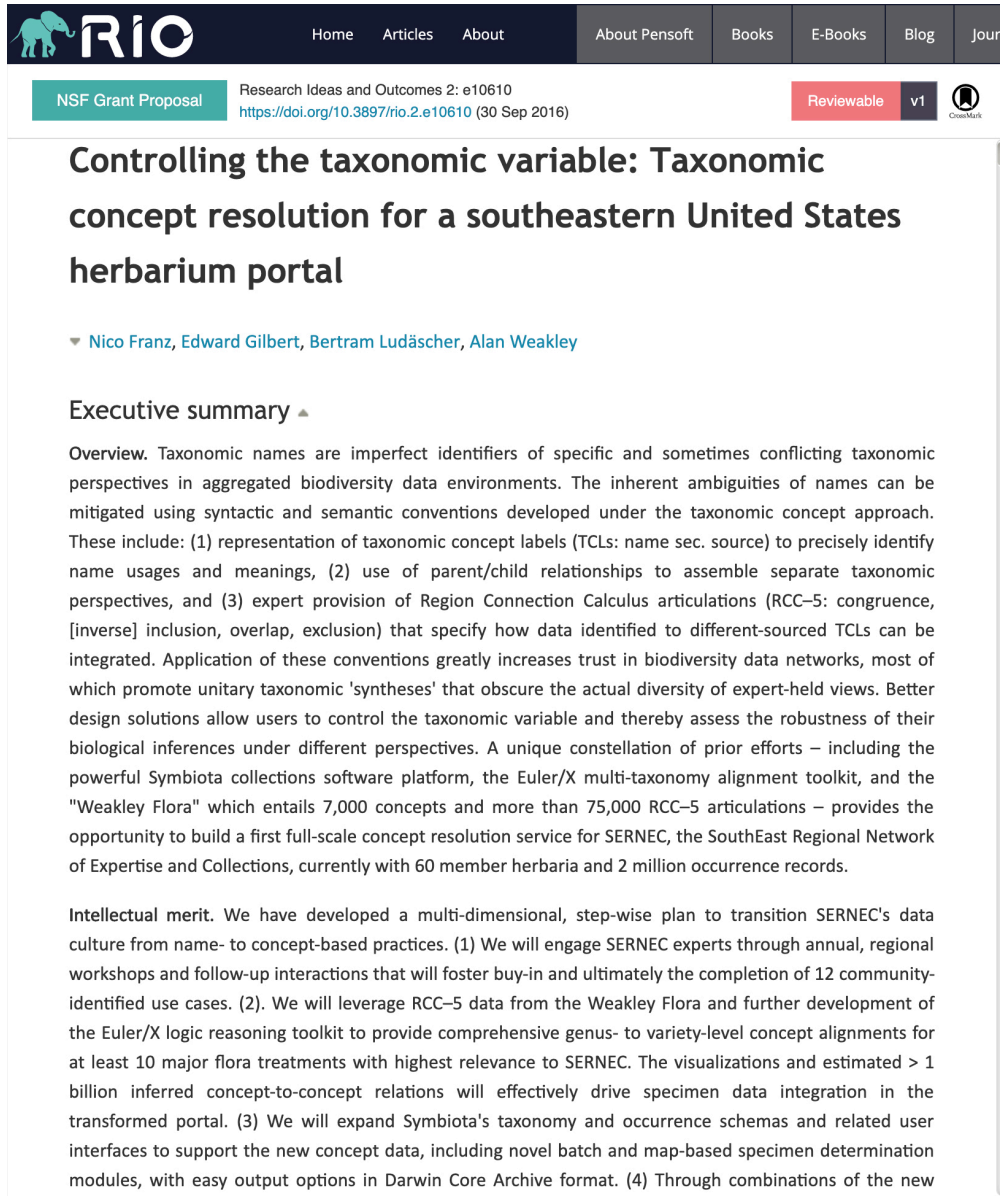
Intellectual merit. We have developed a multi-dimensional, step-wise plan to transition SERNEC's data culture from name- to concept-based practices. (1) We will engage SERNEC experts through annual, regional workshops and follow-up interactions that will foster buy-in and ultimately the completion of 12 community-identified use cases. (2). We will leverage RCC–5 data from the Weakley Flora and further development of the Euler/X logic reasoning toolkit to provide comprehensive genus- to variety-level concept alignments for at least 10 major flora treatments with highest relevance to SERNEC. The visualizations and estimated > 1 billion inferred concept-to-concept relations will effectively drive specimen data integration in the transformed portal. (3) We will expand Symbiota's taxonomy and occurrence schemas and related user interfaces to support the new concept data, including novel batch and map-based specimen determination modules, with easy output options in Darwin Core Archive format. (4) Through combinations of the new

- https://doi.org/10.3897/rio.2.e10610
- Engaging expert/enthusiast communities ⇔ **need for pluralism and democracy** for and **among taxonomic perspectives** in biodiversity data aggregation designs.

Alignment by Alan Weakley (http://herbarium.unc.edu/flora.htm)



Taxonomic data intelligence for Darwin Core occurrences **(3)**



RIO Home Articles About About Pensoft Books E-Books Blog Jour

NSF Grant Proposal

Research Ideas and Outcomes 2: e10610
https://doi.org/10.3897/rio.2.e10610 (30 Sep 2016)

Reviewable v1

Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern United States herbarium portal

Nico Franz, Edward Gilbert, Bertram Ludäscher, Alan Weakley

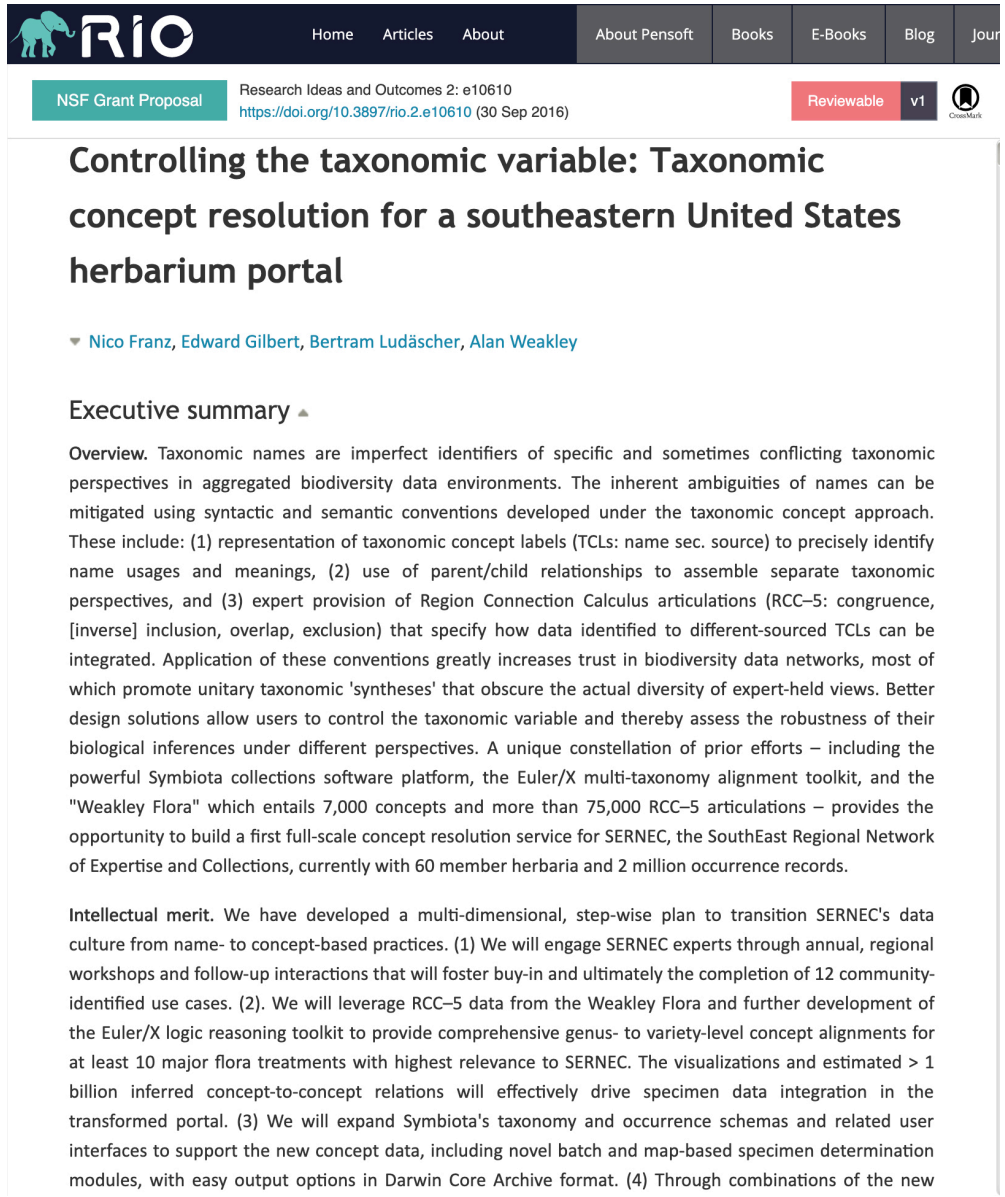
Executive summary

Overview. Taxonomic names are imperfect identifiers of specific and sometimes conflicting taxonomic perspectives in aggregated biodiversity data environments. The inherent ambiguities of names can be mitigated using syntactic and semantic conventions developed under the taxonomic concept approach. These include: (1) representation of taxonomic concept labels (TCLs: name sec. source) to precisely identify name usages and meanings, (2) use of parent/child relationships to assemble separate taxonomic perspectives, and (3) expert provision of Region Connection Calculus articulations (RCC–5: congruence, [inverse] inclusion, overlap, exclusion) that specify how data identified to different-sourced TCLs can be integrated. Application of these conventions greatly increases trust in biodiversity data networks, most of which promote unitary taxonomic 'syntheses' that obscure the actual diversity of expert-held views. Better design solutions allow users to control the taxonomic variable and thereby assess the robustness of their biological inferences under different perspectives. A unique constellation of prior efforts – including the powerful Symbiota collections software platform, the Euler/X multi-taxonomy alignment toolkit, and the "Weakley Flora" which entails 7,000 concepts and more than 75,000 RCC–5 articulations – provides the opportunity to build a first full-scale concept resolution service for SERNEC, the SouthEast Regional Network of Expertise and Collections, currently with 60 member herbaria and 2 million occurrence records.

Intellectual merit. We have developed a multi-dimensional, step-wise plan to transition SERNEC's data culture from name- to concept-based practices. (1) We will engage SERNEC experts through annual, regional workshops and follow-up interactions that will foster buy-in and ultimately the completion of 12 community-identified use cases. (2). We will leverage RCC–5 data from the Weakley Flora and further development of the Euler/X logic reasoning toolkit to provide comprehensive genus- to variety-level concept alignments for at least 10 major flora treatments with highest relevance to SERNEC. The visualizations and estimated > 1 billion inferred concept-to-concept relations will effectively drive specimen data integration in the transformed portal. (3) We will expand Symbiota's taxonomy and occurrence schemas and related user interfaces to support the new concept data, including novel batch and map-based specimen determination modules, with easy output options in Darwin Core Archive format. (4) Through combinations of the new

- https://doi.org/10.3897/rio.2.e10610
- Engaging expert/enthusiast communities ⇔ **need for pluralism and democracy** for and **among taxonomic perspectives** in biodiversity data aggregation designs.
- Spatial reasoning tools (RCC–5) can help attain consistent and comprehensive **taxonomic meaning mappings** ⇔ **intelligence** for data integration across evolving or conflicting views.

Taxonomic data intelligence for Darwin Core occurrences **(3)**



RIO Home Articles About | About Pensoft Books E-Books Blog Jour

NSF Grant Proposal

Research Ideas and Outcomes 2: e10610
https://doi.org/10.3897/rio.2.e10610 (30 Sep 2016)

Reviewable v1

Controlling the taxonomic variable: Taxonomic concept resolution for a southeastern United States herbarium portal

Nico Franz, Edward Gilbert, Bertram Ludäscher, Alan Weakley

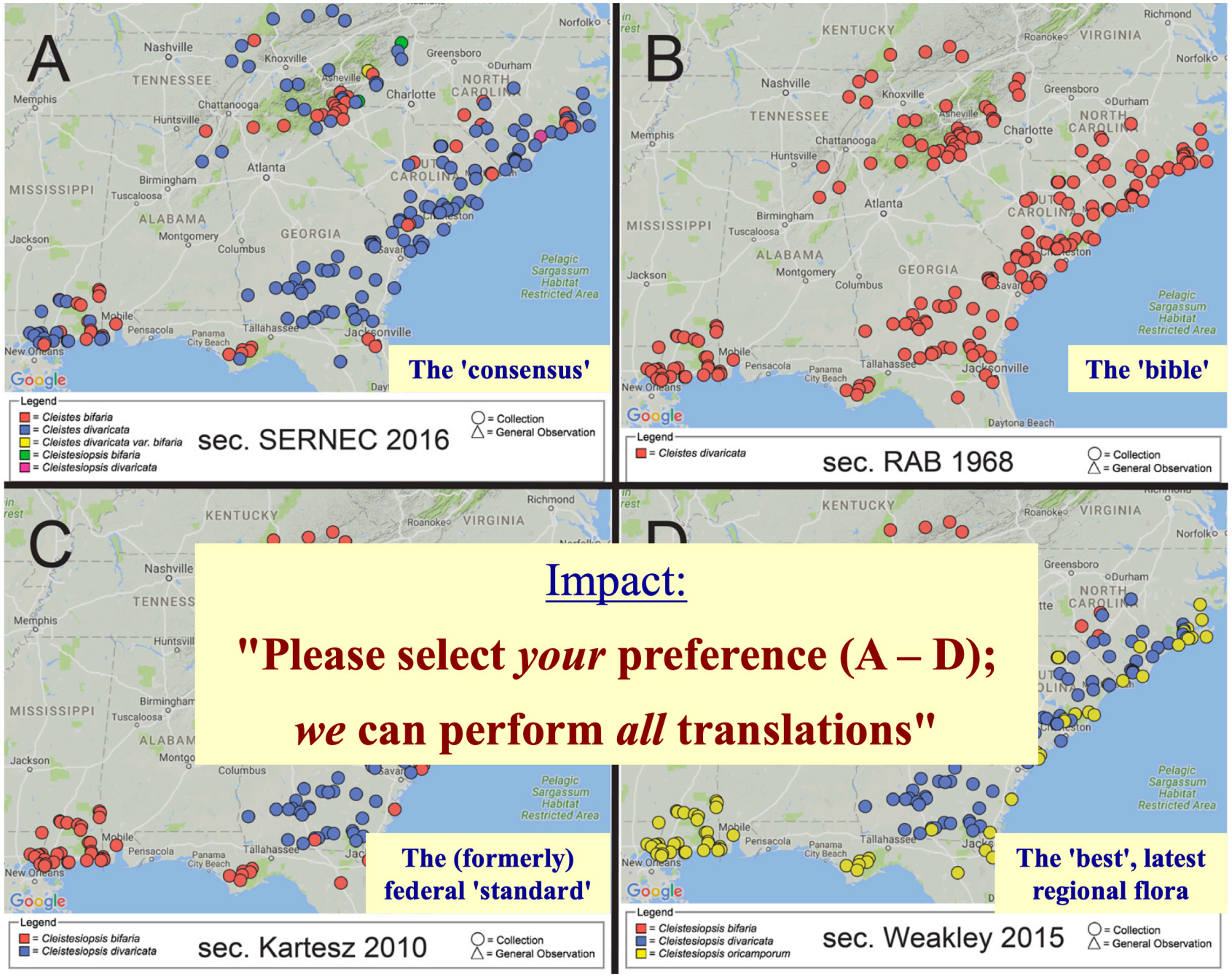
Executive summary

Overview. Taxonomic names are imperfect identifiers of specific and sometimes conflicting taxonomic perspectives in aggregated biodiversity data environments. The inherent ambiguities of names can be mitigated using syntactic and semantic conventions developed under the taxonomic concept approach. These include: (1) representation of taxonomic concept labels (TCLs: name sec. source) to precisely identify name usages and meanings, (2) use of parent/child relationships to assemble separate taxonomic perspectives, and (3) expert provision of Region Connection Calculus articulations (RCC–5: congruence, [inverse] inclusion, overlap, exclusion) that specify how data identified to different-sourced TCLs can be integrated. Application of these conventions greatly increases trust in biodiversity data networks, most of which promote unitary taxonomic 'syntheses' that obscure the actual diversity of expert-held views. Better design solutions allow users to control the taxonomic variable and thereby assess the robustness of their biological inferences under different perspectives. A unique constellation of prior efforts – including the powerful Symbiota collections software platform, the Euler/X multi-taxonomy alignment toolkit, and the "Weakley Flora" which entails 7,000 concepts and more than 75,000 RCC–5 articulations – provides the opportunity to build a first full-scale concept resolution service for SERNEC, the SouthEast Regional Network of Expertise and Collections, currently with 60 member herbaria and 2 million occurrence records.

Intellectual merit. We have developed a multi-dimensional, step-wise plan to transition SERNEC's data culture from name- to concept-based practices. (1) We will engage SERNEC experts through annual, regional workshops and follow-up interactions that will foster buy-in and ultimately the completion of 12 community-identified use cases. (2). We will leverage RCC–5 data from the Weakley Flora and further development of the Euler/X logic reasoning toolkit to provide comprehensive genus- to variety-level concept alignments for at least 10 major flora treatments with highest relevance to SERNEC. The visualizations and estimated > 1 billion inferred concept-to-concept relations will effectively drive specimen data integration in the transformed portal. (3) We will expand Symbiota's taxonomy and occurrence schemas and related user interfaces to support the new concept data, including novel batch and map-based specimen determination modules, with easy output options in Darwin Core Archive format. (4) Through combinations of the new

- https://doi.org/10.3897/rio.2.e10610
- Engaging expert/enthusiast communities ⇔ **need for pluralism and democracy** for and **among taxonomic perspectives** in biodiversity data aggregation designs.
- Spatial reasoning tools (RCC–5) can help attain consistent and comprehensive **taxonomic meaning mappings** ⇔ **intelligence** for data integration across evolving or conflicting views.
- **Biological inferences become robust**

"Controlling the taxonomic variable"



Impact:

"Please select *your* preference (A – D);

***we* can perform *all* translations"**

Source: Franz et al. 2016. Controlling the taxonomic variable: […]. RIO Journal. doi:10.3897/rio.2.e10610

Project remains pending, but look!

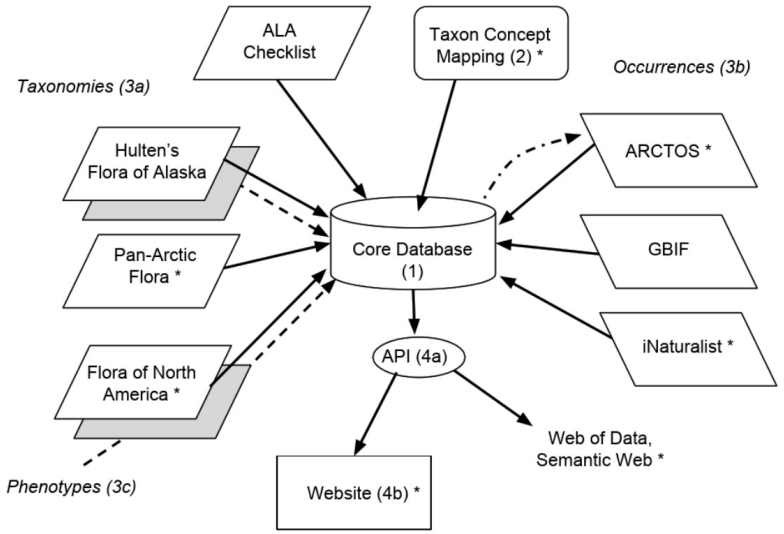
"Taxonomically intelligent data integration for a new **Flora of Alaska**"

Blog 1: Project roadmap

Posted by Cam on 2018-08-17

Hi! Welcome to this development blog, this website, and this project. Steffi and I were so pleased to get this NSF funding — many, many thanks to the ABI panelists, the four reviewers, and to NSF and US taxpayers.

In this blog, I'll lay out the main steps in the development of the informatics infrastructure. (For a general overview of the project, see here.) Here is an overview diagram:



The software and data elements to be constructed and integrated are:

1. **A core database**: a standard **MySQL** (MariaDB) database, running on a web-hosting platform (Dreamhost). Over the years I've played with many database paradigms, including XML databases (ExistDB) and graph databases (e.g., 4store, Allegrograph,

PROJECT INFO

DEV BLOG

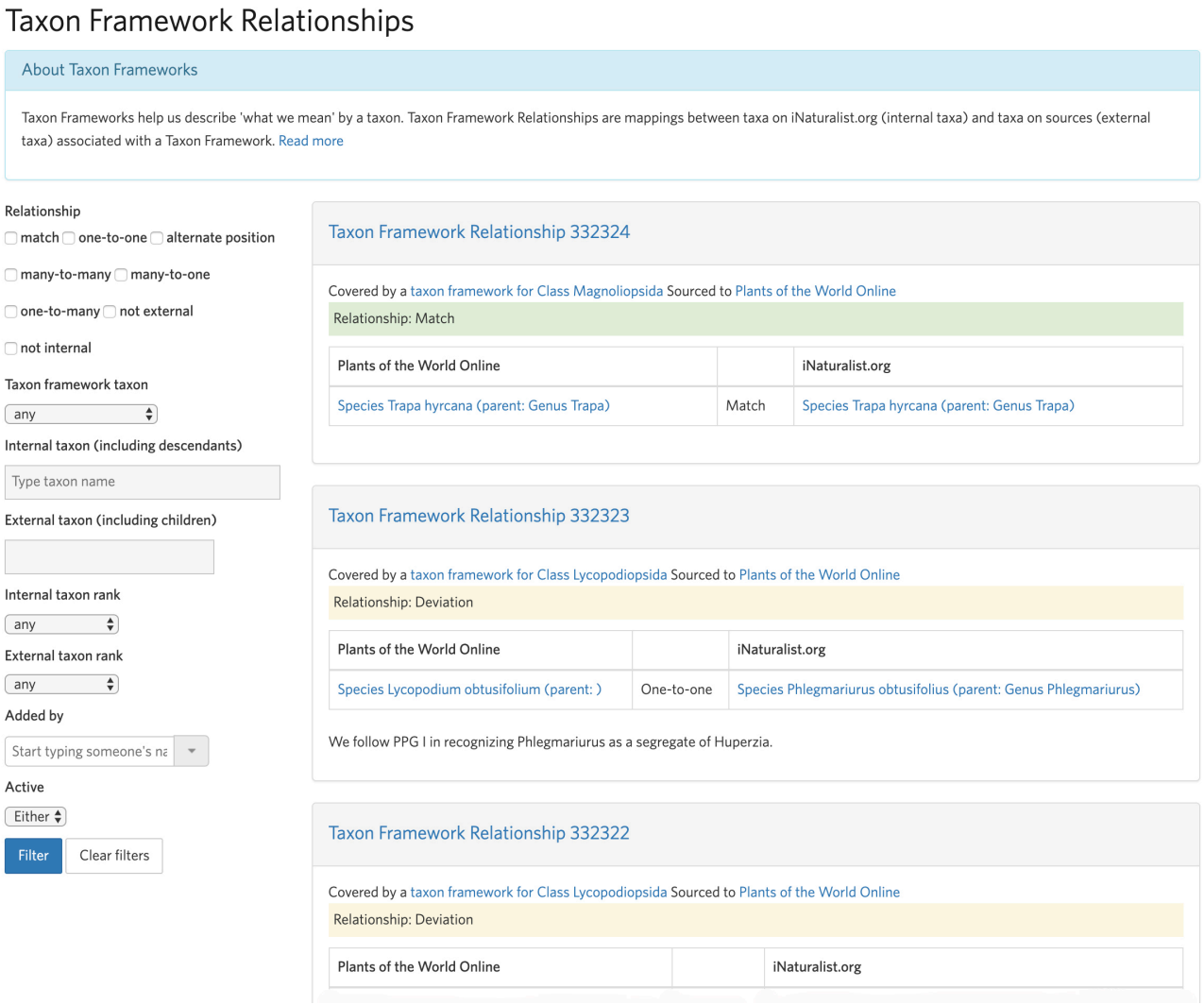
PRODUCTS

PEOPLE

HOME

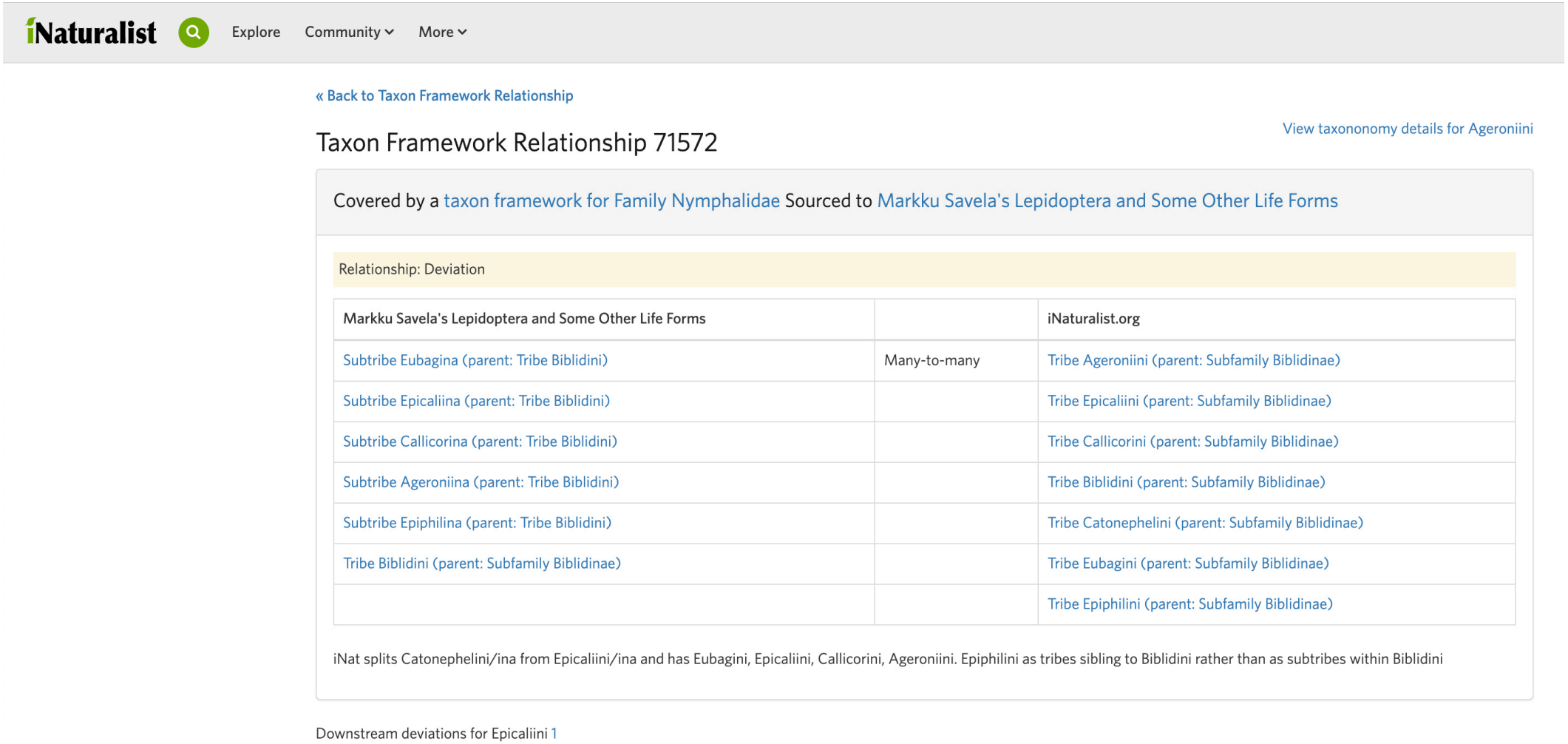
- NSF DBI 1759964
- PIs Ickert-Bond & Webb
- Reconciling Hulten, FNA & Pan-Arctic Flora
- See http://alaskaflora.org/

iNaturalist is mostly there already



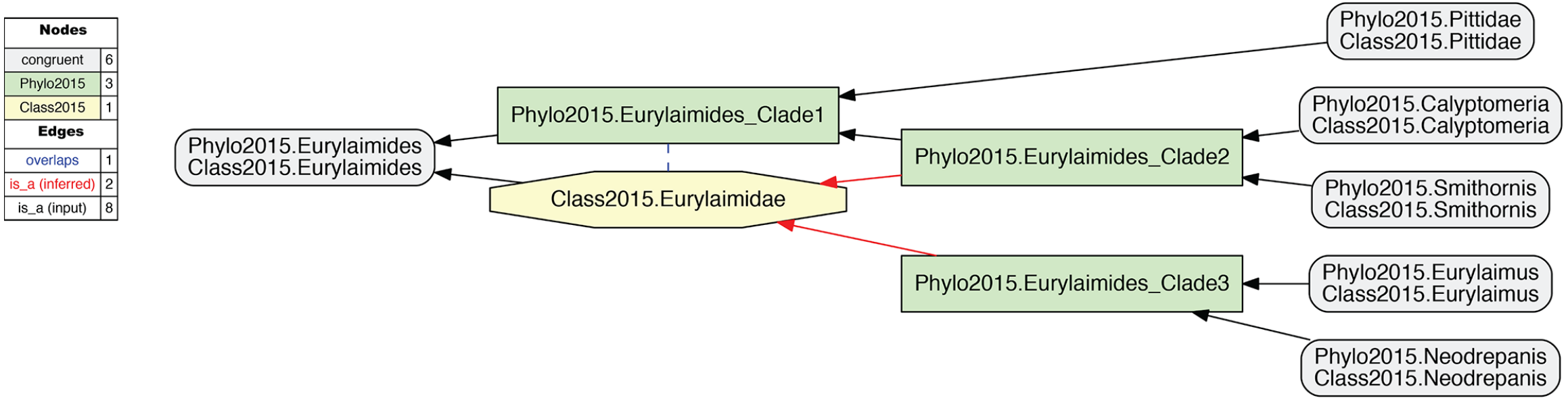
https://www.inaturalist.org/taxon_framework_relationships

iNaturalist is mostly there already



https://www.inaturalist.org/taxon_framework_relationships

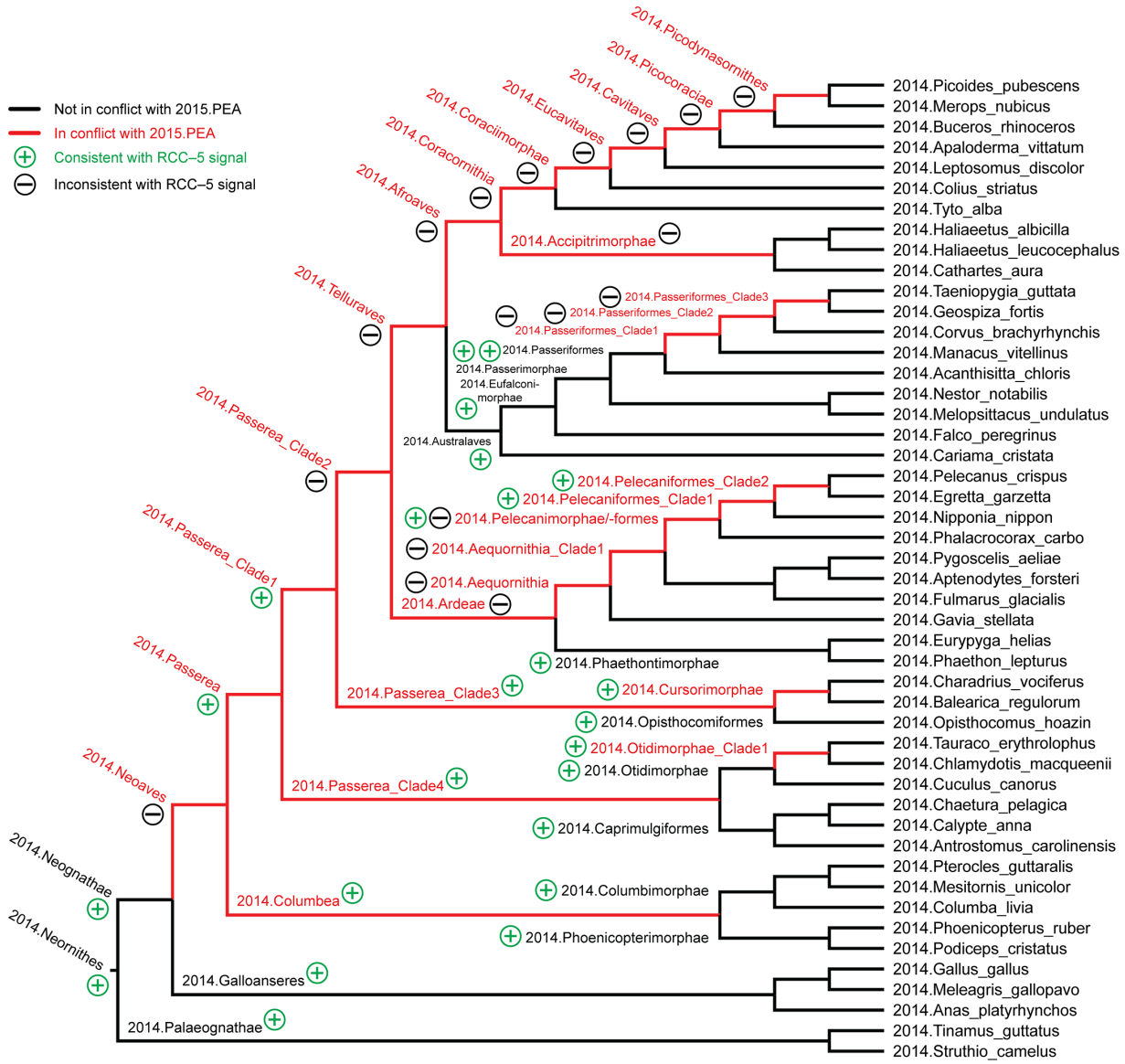
Recent addition: Concept alignment for phylogenomic trees



Franz, N.M. et al. 2019. Verbalizing phylogenomic conflict: […]. https://doi.org/10.1371/journal.pcbi.1006493

Reliable theories of multi-tree node congruence **require expert judgment**

- "Synthesis" OpenTree vs. RCC–5



Franz, N.M. et al. 2019. Verbalizing phylogenomic conflict: […]. https://doi.org/10.1371/journal.pcbi.1006493

Hopeful conclusions

Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

Hopeful conclusions

Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

There are many grassroots or federally supported projects in this domain that **express**

enthusiasm for, and confidence in, a strong future for new, high-volume,

perpetually **data-restructuring systematic research;** driving our evolving

views of relations between DwC datasets and biological knowledge.

Hopeful conclusions

Key message: Biodiversity informatics remains young and fresh; especially if we aim

to **incentivize experts/enthusiasts** in publishing high-quality, "data-intelligent" biodiversity data products.

There are many grassroots or federally supported projects in this domain that **express**

enthusiasm for, and confidence in, a strong future for new, high-volume,

perpetually **data-restructuring systematic research;** driving our evolving

views of relations between DwC datasets and biological knowledge.

If you have the passion and stomach for that future, **join us now!**

Acknowledgments

- Gil Nelson, Alnycea Blackwell, Jillian Goodwin, and all other **iDigBio and Yale Peabody organizers and sponsors of #Digidata 2019.**

- ASU Biocollections & NEON Biorepository team.

- Euler/X team: Bertram Ludäscher, Shizhuo Yu, Jessica Cheng.

- If you wish to read *one* paper on aligning taxonomic concepts: https://doi.org/10.1093/sysbio/syw023

Biodiversity Knowledge Integration Center
Arizona State University