Royal Botanic Gardens Kew

echinopscis:

an extensible notebook for open science on specimens

Nicky Nicolson
Kew Science

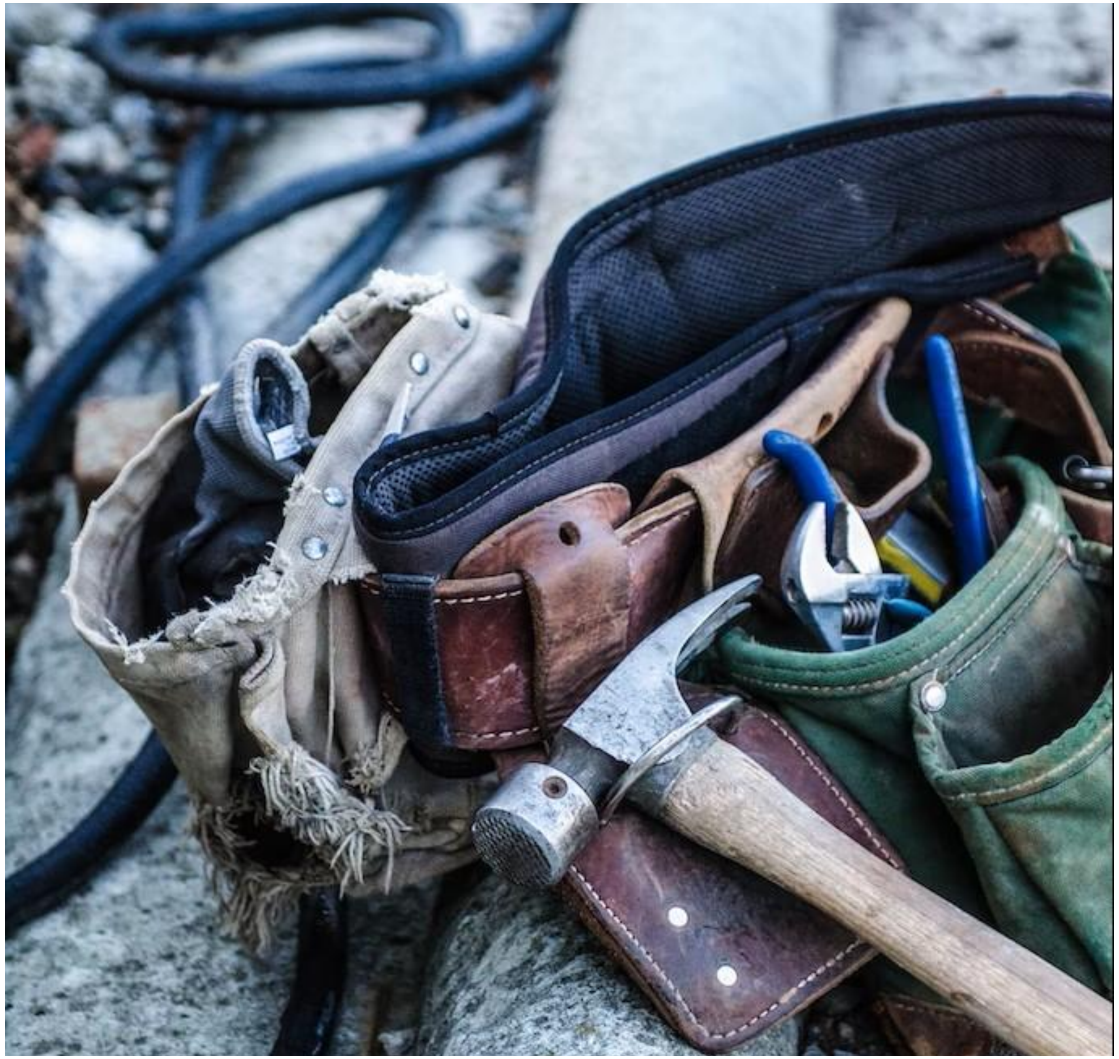
BHL
Biodiversity Heritage Library

#BHLDay2023

Muséum national d'Histoire naturelle (MNHN)
Paris, France, 19 April 2023

Context: personal & institutional

- Transitioned from software development into research
- Open science, take-up & how we design & build for participation
- How we can use software development practices in research:
  - Reuse
  - Automation
  - Version control
  - Dependency management
  - Continuous integration
- Also processes about communication, design & inclusion



BHL launched 2006. Also: Twitter, crowd-sourcing, “e-taxonomy”

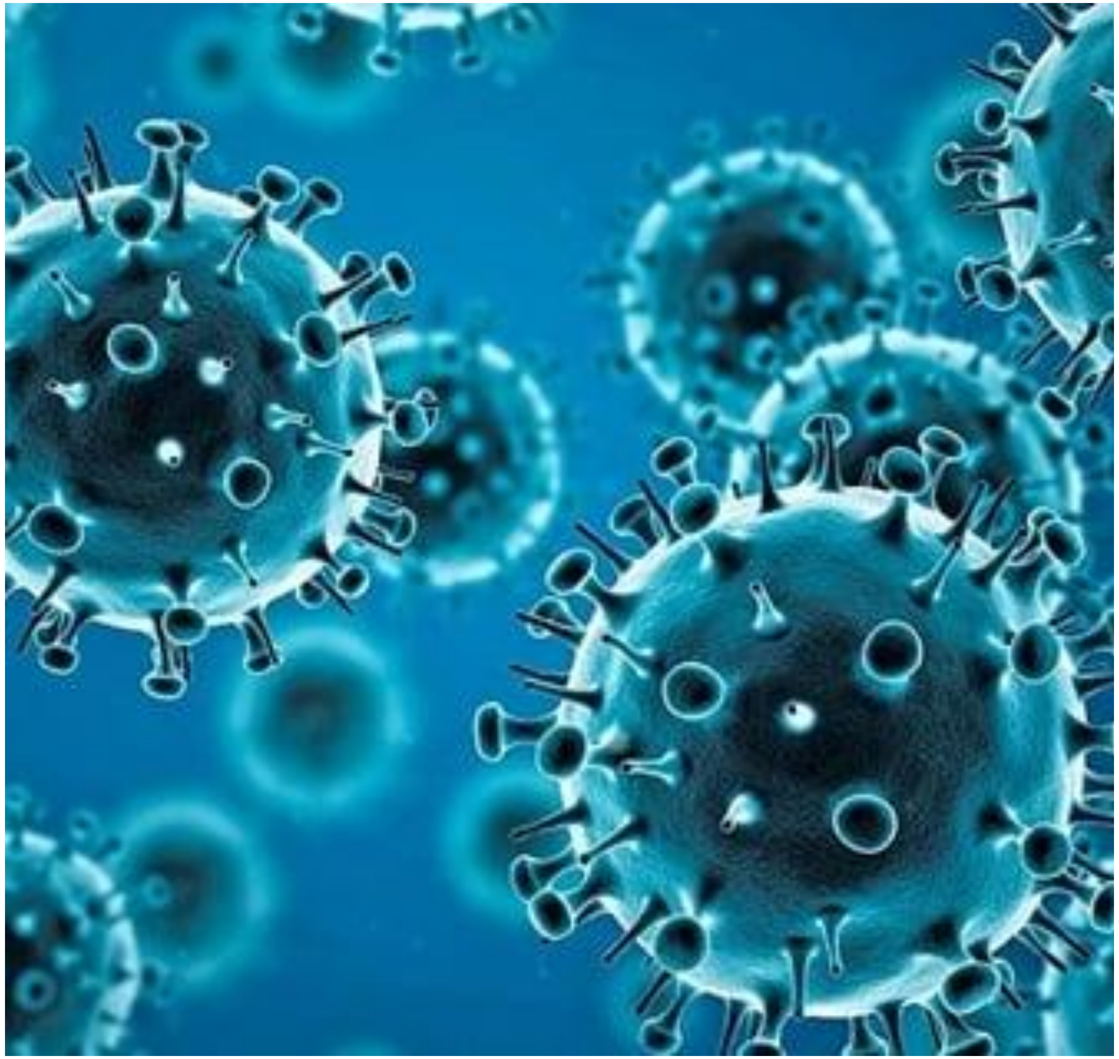
Explosion of data availability online:

- Specimens online (Tracheophyta)
  - 88 million metadata records
  - 38 million images
- Comprehensive taxonomies with distribution
- **Born digital and digitised literature provides context**
- Machine learning / AI: text / images
- Recognised different roles in research
- Bring your idea to the data – compute available online

Where we work: we’re still learning how to transition online

- COVID made us move *everything* online

- When we work online, have we got a usable space, or is it “just another tab”



Who does the work: evolving & inclusive research culture

- Skills development (Carpentries)
- Awareness of different roles in research
    - recognition of the research software engineer role
- Recognition of different activitie required for successful research
- Open science: **open data**
- Online & remote collaboration

PERSISTANT
FINDABLE
MEANINGFUL INTERACTION
ACCESSIBLE
FAIR
PRINCIPLES
INTEROPERABLE
FULL DISCLOSURE
REUSABLE
Scriberia

How we work: building working environments online

- Inclusive design with researchers
- Prototype working environment:
  - Streamline access
  - Manual processes build training data
  - Mobilise rich data to publications & portals
  - A place to plugin AI approaches as these come onstream

OBSIDIAN

- A personal knowledge manager: for creating & linking research notes
- Emphasises linking
- Data stored locally, using open formats
  - Markdown and optional structured data frontmatter
- Works offline
- Extensible architecture – plugins for data access and citation processing
- Active user and developer community

OBSIDIAN

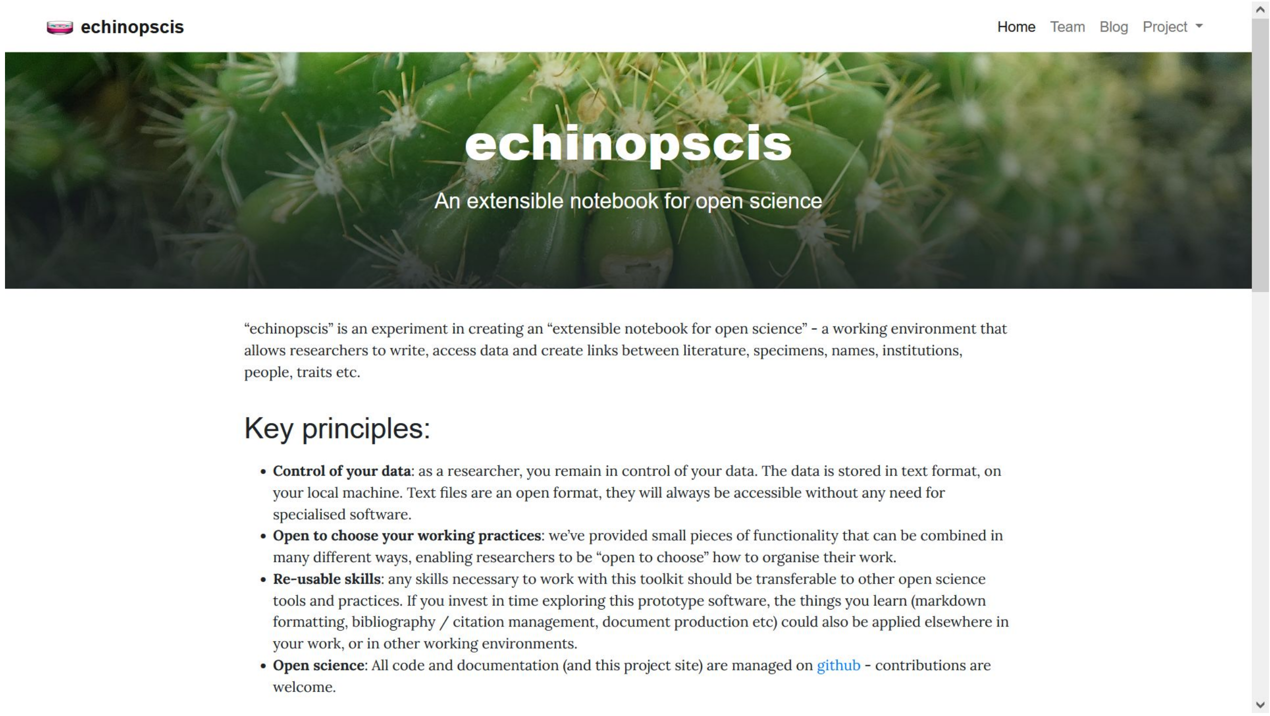
- A personal knowledge manager: for creating & linking research notes
- Emphasises linking
- Data stored locally, using open formats
  - Markdown and optional structured data frontmatter
- Works offline
- Extensible architecture – plugins for data access and citation processing
- Active user and developer community

OpenRefine

Extend Obsidian for specimen research

- Access of relevant data
  - Specimens (GBIF)
  - Names (International Plant Names Index)
  - Collections (Global Registry of Scientific Collections)
  - People (Bionomia)
  - Literature (crossref)
- Creation of links, spatial and network exploration
- Citation in new work
- Open science working practices

echinopscis.github.io



echinopscis

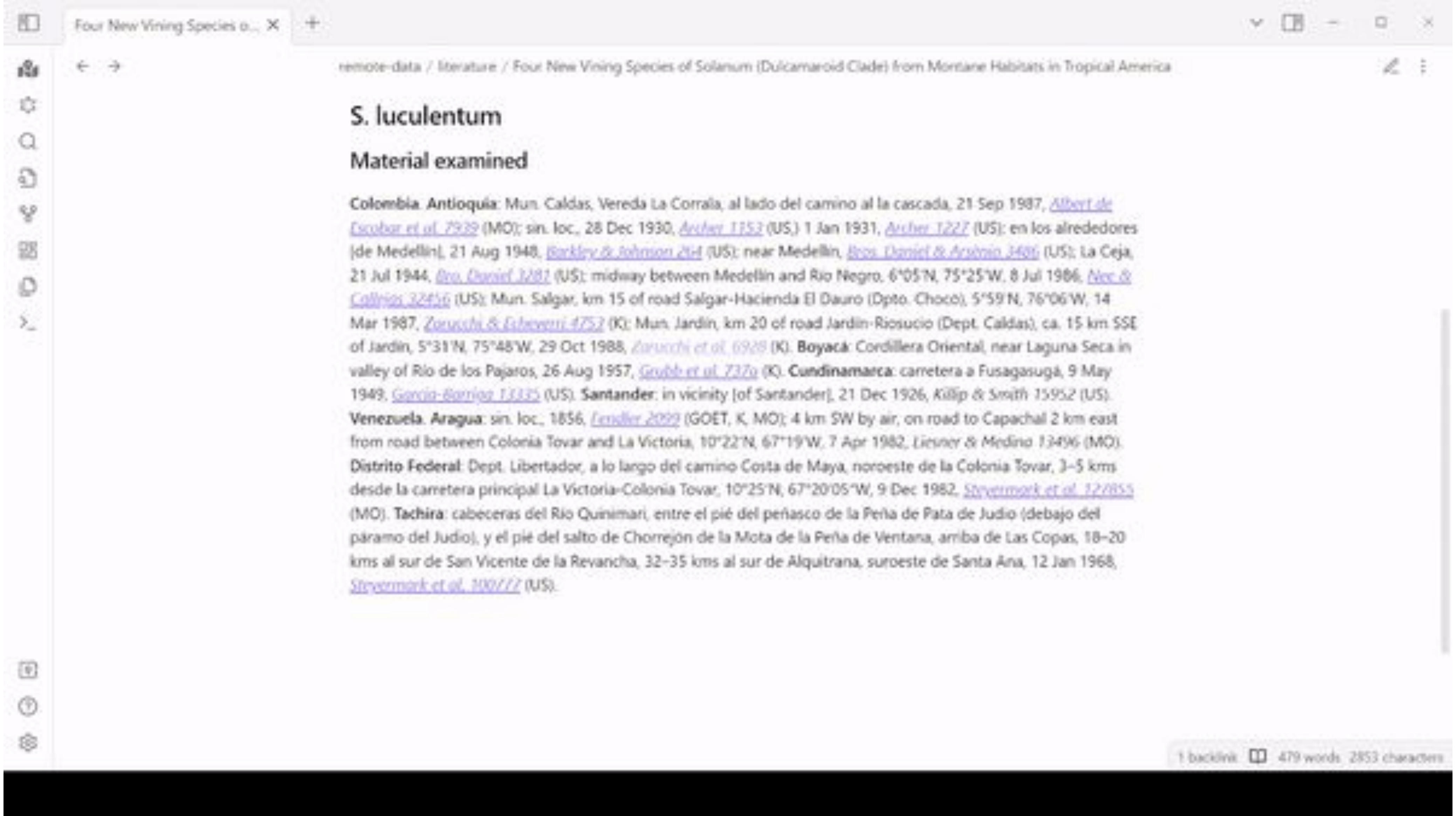
An extensible notebook for open science

“echinopscis” is an experiment in creating an “extensible notebook for open science” - a working environment that allows researchers to write, access data and create links between literature, specimens, names, institutions, people, traits etc.

Key principles:

- **Control of your data**: as a researcher, you remain in control of your data. The data is stored in text format, on your local machine. Text files are an open format, they will always be accessible without any need for specialised software.
- **Open to choose your working practices**: we've provided small pieces of functionality that can be combined in many different ways, enabling researchers to be “open to choose” how to organise their work.
- **Re-usable skills**: any skills necessary to work with this toolkit should be transferable to other open science tools and practices. If you invest in time exploring this prototype software, the things you learn (markdown formatting, bibliography / citation management, document production etc) could also be applied elsewhere in your work, or in other working environments.
- **Open science**: All code and documentation (and this project site) are managed on github - contributions are welcome.

Demo: Unstructured text data to specimen links

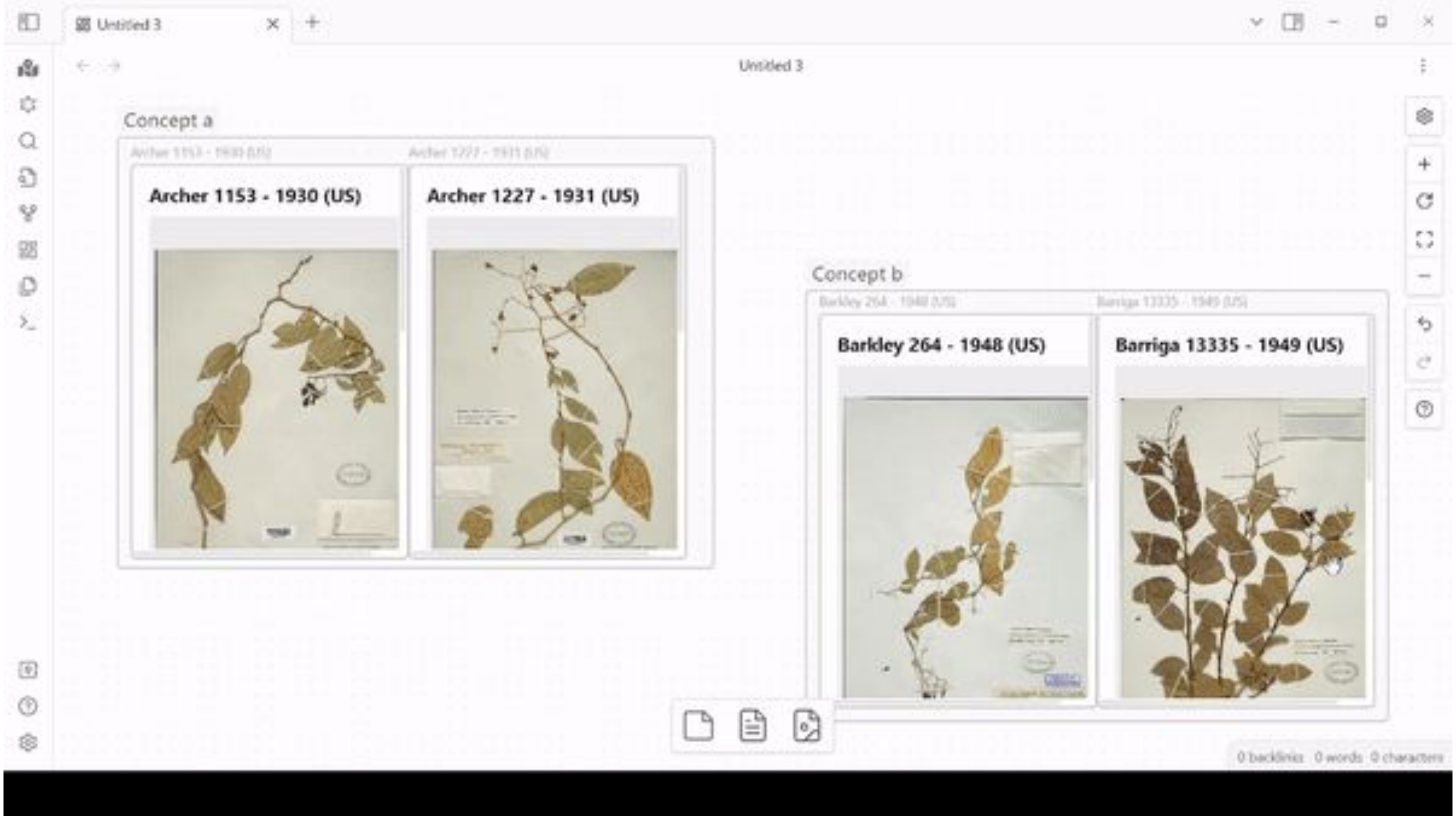

Four New Vining Species o...
remote-data / literature / Four New Vining Species of Solanum (Dulcamaroid Clade) from Montane Habitats in Tropical America
1 backlink
479 words 2853 characters

S. luculentum

Material examined

Colombia. **Antioquia**: Mun. Caldas, Vereda La Corrala, al lado del camino al la cascada, 21 Sep 1987, *Albert de Escobar et al. 7939* (MO); sin. loc., 28 Dec 1930, *Archer 1153* (US.) 1 Jan 1931, *Archer 1227* (US); en los alrededores [de Medellin], 21 Aug 1948, *Barkley & Johnson 264* (US); near Medellin, *Bros. Daniel & Arsênio 3486* (US); La Ceja, 21 Jul 1944, *Bro. Daniel 3281* (US); midway between Medellin and Rio Negro, 6°05'N, 75°25'W, 8 Jul 1986, *Nee & Collings 32456* (US); Mun. Salgar, km 15 of road Salgar-Hacienda El Dauro (Dpto. Choco), 5°59'N, 76°06'W, 14 Mar 1987, *Zarucchi & Echeverri 4753* (K); Mun. Jardin, km 20 of road Jardin-Riosucio (Dept. Caldas), ca. 15 km SSE of Jardin, 5°31'N, 75°48'W, 29 Oct 1988, *Zarucchi et al. 6928* (K). **Boyacá**: Cordillera Oriental, near Laguna Seca in valley of Rio de los Pajaros, 26 Aug 1957, *Grubb et al. 737a* (K). **Cundinamarca**: carretera a Fusagasugá, 9 May 1949, *Garcia-Barriga 13335* (US). **Santander**: in vicinity [of Santander], 21 Dec 1926, *Killip & Smith 15952* (US). **Venezuela**. **Aragua**: sin. loc., 1856, *Fendler 2099* (GOET, K, MO); 4 km SW by air, on road to Capachal 2 km east from road between Colonia Tovar and La Victoria, 10°22'N, 67°19'W, 7 Apr 1982, *Liesner & Medina 13496* (MO). **Distrito Federal**: Dept. Libertador, a lo largo del camino Costa de Maya, noroeste de la Colonia Tovar, 3–5 kms desde la carretera principal La Victoria-Colonia Tovar, 10°25'N, 67°20'05"W, 9 Dec 1982, *Steyermark et al. 127855* (MO). **Tachira**: cabeceras del Rio Quinimari, entre el pié del peñasco de la Peña de Pata de Judio (debajo del páramo del Judio), y el pié del salto de Chorrejon de la Mota de la Peña de Ventana, arriba de Las Copas, 18–20 kms al sur de San Vicente de la Revancha, 32–35 kms al sur de Alquitrana, suroeste de Santa Ana, 12 Jan 1968, *Steyermark et al. 100777* (US).

Demo: “Workbench” specimen examination



Provisional roadmap (open to influence)

1. **Personal research environment** based on Markdown authoring and linking
2. **Web publication** using static site generators (conceptual similarities with the GBIF hosted portal work)
3. **Document production:** with structured bibliographic/specimen references
4. **Dataset production:** mobilisation of content and links into DarwinCore archives for aggregator harvesting

BHL relevance (1) data: integration of “meta” works like TL-2

- **TL-2 is entity rich:**
  - People
  - Places
  - Institutions
  - Timelines
  - Expeditions
  - Bibliographic works
  - Eponyms
- We have corrected text data in addition to BHL OCR (Smithsonian data package)
- Reference resource for researchers working on Bionomia

mobilise

Bionomia

Link natural history specimens to the world's collectors

BHL relevance (2) tools: provision of reference parsing APIs

Royal Botanic Gardens Kew

- echinopscis relies on “strings to things” style APIs - pass a piece of text like might be found in unstructured paragraph, get a reference to an entity back
- BHL API requires atomised input
- Role for BHL to provide reference parsing (whatever the approach, BHL can learn from usage traffic):
    - anystyle.io
    - ChatGPT

BHL relevance (3) approach: community development

Promoting use of generic tools through skills development, investing in development of research culture

- **Library Carpentry** (short training courses)
  - Use of reference managers
  - Automated bibliographic production
  - Article & author identification (DOI / ORCID)
  - Deposition
- **Open Life Sciences** (16 week mentoring):
  - Project management: “open by design”
  - Community building, inclusivity

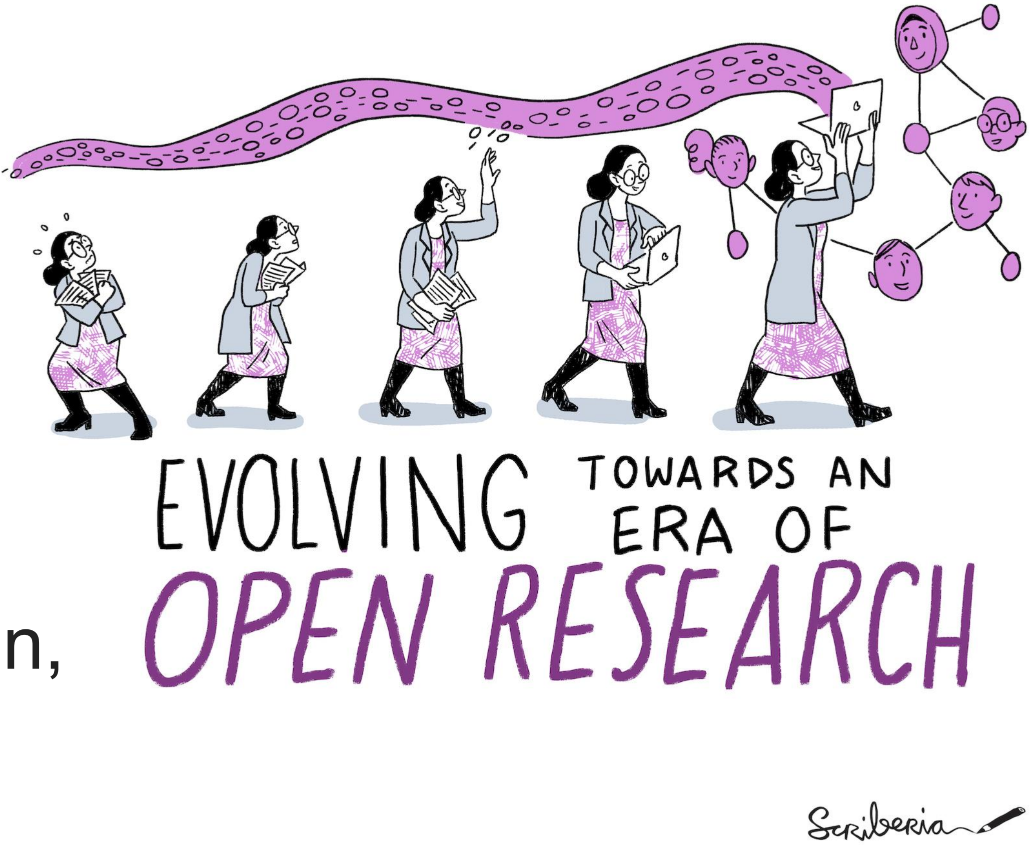
Conclusions: “e-taxonomy” lives(!) as “open science”

- We’ve built an open project based on participation & reusable skills

- Include people where they are now, show a relevant path to destination

- Extensible with AI approaches (entity identification, clustering, summarisation, link prediction)

- Usable today with generic skills - try it out: https://echinopscis.github.io



n.nicolson@kew.org / @nickynicolson / @nickynicolson@mastodon.social

Image credits

- Slide 2: Leather toolbelt by jesse orrico (jessedo81) on Unsplash
- Slides 3 & 6: RBG Kew
- Slide 4: WHO
- Slides 5 & 16: The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807

Useful links

- echinopscis: https://echinopscis.github.io
- Obsidian: https://obsidian.md/ & Obsidian roundup (weekly newsletter): https://www.eleanorkonik.com/tag/roundup/
- Open Refine: https://www.openrefine.org
- Training resources:
  - Carpentries (including Library Carpentry): https://carpentries.org/
  - The Turing Way: https://the-turing-way.netlify.app/
  - Open Life Sciences: https://openlifesci.org/

Contact

- Email: n.nicolson@kew.org
- Github: @nickynicolson
- Twitter: @nickynicolson
- Mastodon: @nickynicolson@mastodon.social
- Web: https://www.kew.org/science/our-science/people/nicky-nicolson