Specimens as research objects: reconciliation across distributed repositories to enable metadata propagation

Nicky Nicolson[1,3], Alan Paton[2], Sarah Phillips[2], Allan Tucker[3]

(1) Biodiversity Informatics & Spatial Analysis, RBG Kew (UK), (2) Collections, RBG Kew (UK), (3) Department of Computer Science, Brunel University London (UK).

n.nicolson@kew.org @nickynicolson

IEEE eScience, Amsterdam (NL), 29th October - 1st November 2018.

Royal Botanic Gardens Kew

14th eScience IEEE INTERNATIONAL CONFERENCE

Brunel University London

Structure

- Background: Specimens in a digital world
- Manually reconciling specimens using digitised data
- Institutional scale reconciliation: using data mining
- Applications:
  - annotation propagation
  - data-derived institutional network
- Conclusions

Background: Specimens in a digital world

Systematics as escience

Systematics

- Naming & classifying organisms
- Shared physical specimens
- Annotated in repositories, referenced in literature

Recently:

- Specimen (and literature) digitisation
- Computable data repurposed for modelling, data mining
- Potential for data flows based on shared, linked digital objects

eScience rationale & scope

"offers a platform for digital technologies to advance research - from the humanities to the physical sciences."

"...The aim of this conference is cross-fertilization across academic disciplines, to advance academic research by fully exploiting the use of digital technologies."

Open science: Sharing research outputs, throughout process

"Open science" ~ systematics

Open science: share research outputs to aid understanding & reproducibility

Collect multiple specimens

Distribute to open repositories

Digitisation challenge

Scale of "problem":

- 400 million specimens in 3000 collections
- Requires huge effort
- Multiple,distributed, overlapping projects

Useful attributes of specimens:

- Flat, digitised easily
- Large, annotable: data rich

Aggregation of digital data:

- Data standards enable aggregation
- Aggregation ramps up re-use
  - **and** re-purposing to new uses

www.gbif.org
GLOBAL BIODIVERSITY INFORMATION FACILITY

Digitisation example

Specimen imaged: 🖼

Label data transcribed: ▦

The "what / where / when" data

UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11

Two products: image and structured data. Presented using data standards, shared with aggregators

Manually reconciling specimens using digitised data

Why reconcile?

Specimens are research objects, annotated and cited as evidence, with established citation convention.

Cited in literature:

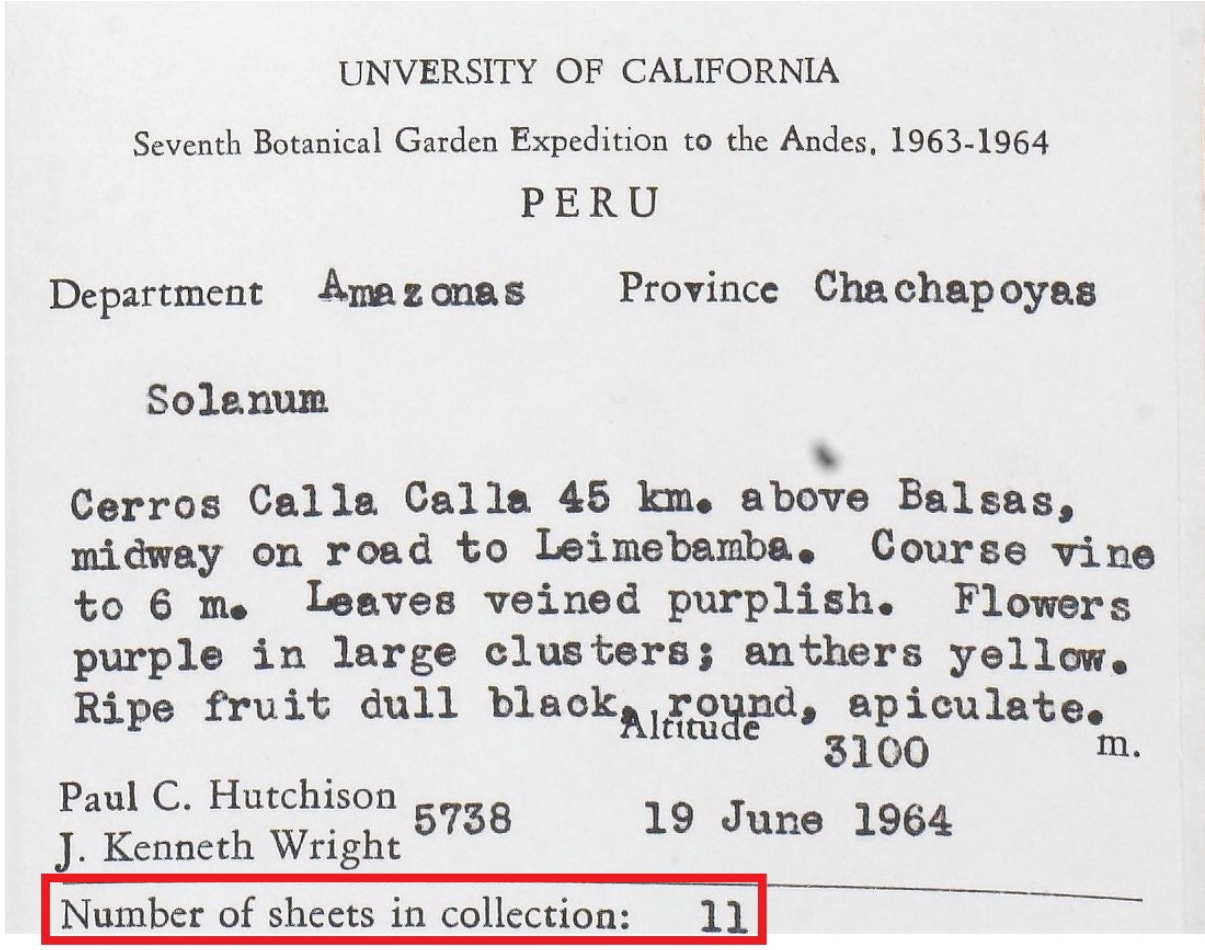
> **Additional specimens examined. Peru**. Amazonas: Balsas-Leimebamba road, km 406, 4 Jun 1977, Boeke 1927 (MO); Prov. Chachapoyas, 29 Jul 1991, Mostacero et al. 2619 (MO); Prov. Chachapoyas, Atuén, Chuquibamba, 18 Jul 1995, Quipuscoa & Bardales 187 (BM, F, MO); middle eastern slopes, near kms 411–416 of Leimebamba-Balsas road, 11 Jul 1962, Wurdack 1314 (K, USM).

doi:10.1371/journal.pone.0010502

Citation components:

- **Date**: 11 Jul 1962
- **Collector name**: Wurdack
- **Collectors number**: 1314
- **Institution codes**: K, USM

Specimen group: Hutchison 5738



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11

Manual reconciliation process

1. Look through digitised records to find specimens generated from the collection event labelled Hutchison 5738
2. Separate metadata items into three distinct categories:
    1. Collection event data
    2. Institution codes - who holds the specimens generated from the collection event
    3. Metadata generated post-accession - determinations, georeferences etc.

Specimen group: Hutchison 5738

Collection event			Holder	Post-accession data					
recorded by	record number	event date	institution code	scientific name	referenced in article	digitised	managed as type	georeferenced	imaged
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum aligerum Schltdl.	-	✔	-	-	-
Hutchison, P.C.	5738	1964-06-19	K	Solanum sanchez-vegae S.Knapp	✔	✔	✔	✔	✔
Paul C. Hutchison—J. Kenneth Wright	Hutchison 5738	1964-06-19	MO	Solanum cutervanum Zahlbr.	-	✔	-	-	-
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	P	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-
P. C. Hutchison & J. K. Wright	5738	1964-06-19	US	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	USM	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-

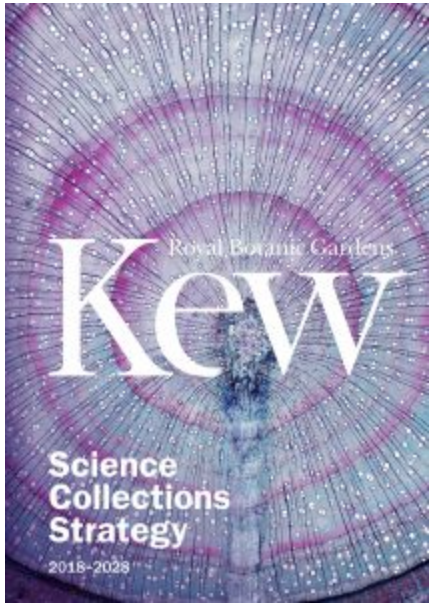
Specimen group: Hutchison 5738

Collection event			Holder	Post-accession data					
recorded by	record number	event date	institution code	scientific name	referenced in article	digitised	managed as type	georeferenced	imaged
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum aligerum Schltdl.	-	✔	-	-	-
Hutchison, P.C.	5738	1964-06-19	K	Solanum sanchez-vegae S.Knapp	✔	✔	✔	✔	✔
Paul C. Hutchison—J. Kenneth Wright	Hutchison 5738	1964-06-19	MO	Solanum cutervanum Zahlbr.	-	✔	-	-	-
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	P	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-
P. C. Hutchison & J. K. Wright	5738	1964-06-19	US	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	USM	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-

Specimen group: Hutchison 5738

Collection event			Holder	Post-accession data					
recorded by	record number	event date	institution code	scientific name	referenced in article	digitised	managed as type	georeferenced	imaged
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum aligerum Schltdl.	-	✔	-	-	-
Hutchison, P.C.	5738	1964-06-19	K	Solanum sanchez-vegae S.Knapp	✔	✔	✔	✔	✔
Paul C. Hutchison—J. Kenneth Wright	Hutchison 5738	1964-06-19	MO	Solanum cutervanum Zahlbr.	-	✔	-	-	-
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	P	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-
P. C. Hutchison & J. K. Wright	5738	1964-06-19	US	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	USM	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-

Can we do this at institutional scale?



Enhancing the analysis of collection data (p. 49):

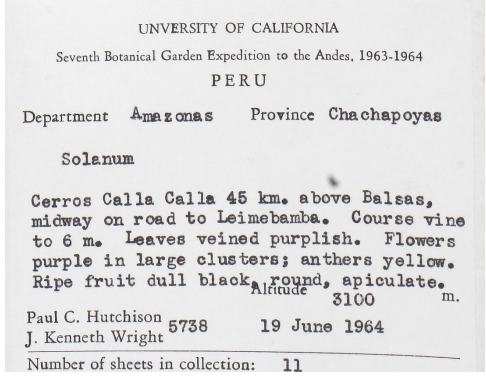
...**we will embrace new machine learning techniques to enhance the curation and analysis of the Science Collections**. Possible applications of these techniques to Kew's collections include:

- **accelerating the interconnection of collection data**
- **institutional-scale analysis of collections**
- **comparison with other institutional collections**

Institutional scale reconciliation: using data mining

"What, where, when" to "Who & why"

Structured data



UNVERSITY OF CALIFORNIA

Seventh Botanical Garden Expedition to the Andes, 1963-1964

PERU

Department Amazonas Province Chachapoyas

Solanum

Cerros Calla Calla 45 km. above Balsas, midway on road to Leimebamba. Course vine to 6 m. Leaves veined purplish. Flowers purple in large clusters; anthers yellow. Ripe fruit dull black, round, apiculate.

Altitude 3100 m.

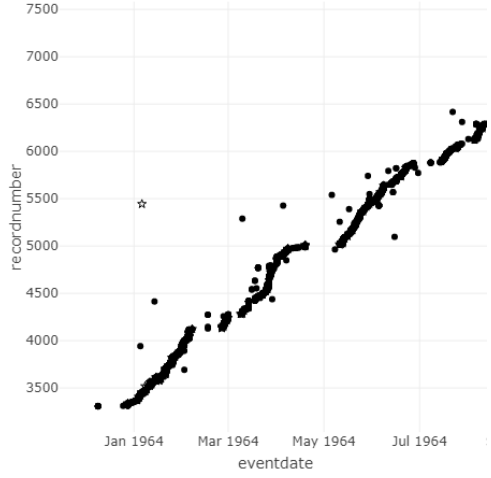
Paul C. Hutchison
J. Kenneth Wright 5738 19 June 1964

Number of sheets in collection: 11

Numeric attributes:

- What? Collector's recordnumber
- When? Event date

Positive correlation



Data mining process based on DBSCAN clustering allocates specimens to collectors based on their presence in collector "traces"

N. Nicolson and A. Tucker, "Identifying Novel Features from Specimen Data for the Prediction of Valuable Collection Trips," presented at the International Symposium on Intelligent Data Analysis, 2017, pp. 235–246. doi:10.1007/978-3-319-68765-0_20

Using a data mined collector ID

Easier to query the data:

- Show me all specimens collected by X

Different way to group up the data:

- Traditionally we detect duplicates using collector name, number and year
- Simple to find duplicates:
  - Collector ID
  - Collectors record number
  - Date

Share any annotations and allied data created post-accession amongst specimen duplicate holders

- Images
- Georeferences
- Determinations
- Type citations

Reconciliation

Group on data mined collector ID, event date and recordnumber

Assessment

Examine variation in countrycode, and higher taxonomy amongst candidate groups.

Conservative approach: only those groups with no variation taken forward for further analysis.

Results

7,102,710 specimens assessed to participate in duplication relationships, of an input set of 19,827,998.

Applications: annotation propagation

Pass georeferences to holders

Collection event			Holder	Post-accession data					
recorded by	record number	event date	institution code	scientific name	referenced in article	digitised	managed as type	georeferenced	imaged
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P. C. Hutchison & J. K. Wright	5738	1964-06-19	F	Solanum aligerum Schltdl.	-	✔	-	-	-
Hutchison, P.C.	5738	1964-06-19	K	Solanum sanchez-vegae S.Knapp	✔	✔	✔	✔	✔
Paul C. Hutchison—J. Kenneth Wright	Hutchison 5738	1964-06-19	MO	Solanum cutervanum Zahlbr.	-	✔	-	-	-
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P. C. Hutchison	5738	1964-06-19	NY	Solanum sanchez-vegae S.Knapp	-	✔	✔	✔	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	P	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-
P. C. Hutchison & J. K. Wright	5738	1964-06-19	US	Solanum sanchez-vegae S.Knapp	✔	✔	✔	-	✔
P.C. Hutchison & J.K. Wright	5738	1964-06-19	USM	Solanum sanchez-vegae S.Knapp	✔	-	?	-	-

Detection

Examine groups, look for uneven population of annotations (a mix of set and unset).

Count how many specimens could receive updates.

Results

- Type citations: 93,044
- Georeferences: 1,121,865
- Images: 1,097,168
- Scientific names: 2,191,179 specimens exist in groups holding multiple scientific names.

Applications: data-derived institutional network

Institution codes part of specimen duplicate groups

A duplicate group is a collection of specimens.

Contains a list of the institutional holders of component specimens.

From the Hutchison 5738 example:

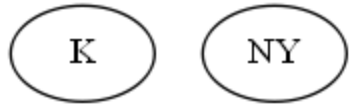
Institution codes: (F, K, MO, NY, P, US, USM)

Question:

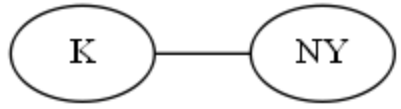
Do we see the same institution codes co-occurring across many different specimen duplicate groups?

Reshape specimen groups to graph

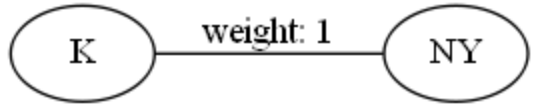
Institution codes are the nodes



Create a relationship if two institution codes co-occur in a specimen duplicate group



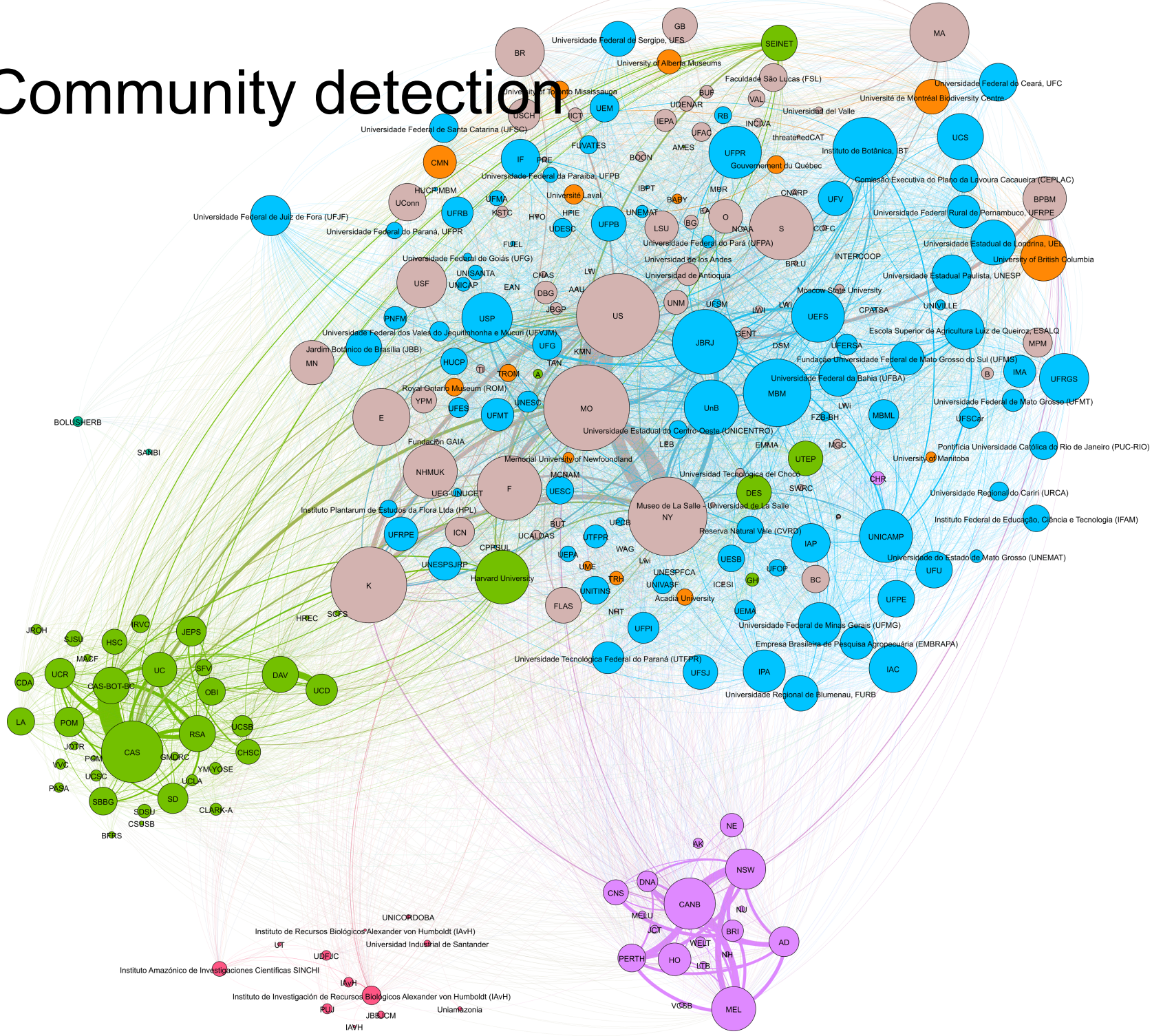
Weight the relationship: incrementing weight every time we see a co-occurrence



Graph metrics: 260 nodes, 6,588 weighted edges.

Graph analysis: Louvain community detection.

Community detection



Conclusions

Future work

Process refinement

- Singleton analysis - likely very few true singletons, but sibling specimens not currently digitally available
- More precise estimates about the undigitised dataset

Refining network analysis

- Adding directionality to the inferred network
- Examining the direction of annotation flow

Implications of results

- Research recognition of some annotation types

Conclusions

- Specimens are research objects managed for long term consultation & annotation in their physical form
- Specimens form a shared global resource
- Reconciling duplicates shares effort, overcomes fragmented information management
- Possible to build a data-derived institutional network: demonstrates existing (& potential) co-working

Specimens as research objects: reconciliation across distributed repositories to enable metadata propagation.

N. Nicolson, A. Paton, S. Phillips & A. Tucker, IEEE eScience, Amsterdam (NL), 29th October - 1st November 2018.

n.nicolson@kew.org @nickynicolson

Paper: doi:10.1109/eScience.2018.00028 / e-print: arXiv:1809.07725

Slides: doi:10.6084/m9.figshare.7327325

Acknowledgements

Data providers: for sharing their specimen data using open standards.

Photo credits: W.J.Baker (field collections), A.Monro (Kew herbarium).