WORKING GROUP 2 REPORT

Sofia, BU, 12-13 March 2019

WG 2 WORKSHOP Authority management for people names

1. Attendance:

   a. WG Leader and Co-Leader: Elspeth Haston (RBGE) and Arnald Marcer (CREAF)

   b. Convener/Rapporteur: Elspeth Haston (RBGE), Quentin Groom (MBG) and Arnald Marcer (CREAF)

   c. Number of Participants: 17

   d. Inclusiveness rate of the WG participation:

   - Gender balance (Female: 5; Male: 12)
   - Geographical distribution (AT: 1; BE: 1; DE: 1; ES: 1; FR: 3; BU: 2; UK: 5; Australia: 1; Canada: 1; USA: 1)
   - Seniority: at least 2 people were Early Career Researchers

2. Objectives of the workshop

   The drive to digitise the natural history collections of the world has the potential to link massive amounts of biodiversity and cultural data across time and place. However, the level to which this will enable the cross-discipline research and analysis that we aspire to is dependent on being able to make connections effectively between the data. The long-standing issues of linking biodiversity data through people, such as the collector, determiner or author have still not been fully resolved, although innovative approaches are being proposed and tested. Options already exist for the authoritative identification of people, however, they differ in their implementation and management, which means we must consider their suitability for use with natural history collections. Importantly, any system will need to enable links to contemporary and historic resources from all sciences and cultural disciplines.

   This workshop brought together key contributors from across the relevant range of disciplines to agree a strategy for adoption and implementation for the authority management of people names in natural history collections and cross-discipline research and exposure.

The overall objectives of the workshop were:

- To agree on clearer systems for the attribution of units of research work (collecting, determining) not currently covered by traditional publication focussed career credit metrics
- To determine what can be facilitated by better connections btw data
    - better understanding of biases in the data
    - more opportunities for shared curation
    - improved quality control (more resources for georeferencing etc)
    - enabling different kinds of research (history of science, analyses of participation etc)

3. Topics discussed

- In reference to completion of the objectives of year one

Objective 2: Bridging: Linking complementary expertise of information scientists, biodiversity researchers and geoscientists leading to new concepts, technical innovations and products and Task a) identifying existing and new communities and expertise.

Objective 3: Compiling: Develop recommendations and best practices linking regional and global community standards and guidelines.

1. Review of current expertise in people name identifiers.
2. Review of existing systems for managing people names.
3. Assessment of pilot case studies showing the implementation of people name identifiers within collections.
4. Identification of major issues and challenges.
5. Agreement on a system of identifiers for people names based on several existing platforms.

- Towards the definition of activities and actions for year two.

1. Agreement to hold a pre-conference workshop and a symposium at the Biodiversity_Next conference in October in Leiden.
2. Additional institutes will aim to implement and test the existing pilot case studies.
3. Institutes will assess their existing people name data structure and, where possible, add the agreed identifiers.

4. Major outcomes

The workshop brought together a group of experts from a range of organisations and disciplines, including industry partners, to create an extended network of people to tackle a major issue in the digitisation process – that of assigning the correct person to the event related to the specimen. This is one of the tasks that can both slow down the digitisation workflow as well as potentially causing problems in data quality.

A review of existing people name identifier systems was carried out. Existing pilot case studies were presented. The major issues and challenges were identified and collated. A system of identifiers based on several existing platforms was proposed and recommended. This would include ORCID, Wikidata and ISNI.

The progress achieved during the workshop will continue in year two, with a pre-conference workshop and a symposium at the Biodiversity_Next conference in October in Leiden. During the summer, additional institutes will aim to implement and test the existing pilot case studies.

5. Challenges faced for year two

    a. Maintaining the momentum from the workshop will be important to achieve the progress planned for year two.
    b. Ensuring that the communication channels work effectively.
    c. The brokerage of multiple identifiers.
    d. Agreeing on a methodology for teams of people, *e.g.* collectors
    e. Ensuring that the system can manage the order of names.
    f. The process of disambiguation for names.
    g. The needs and motivations of the authors, collectors and institutions.
    h. Ownership of your ID (for the living).
    i. Identifiers for living collectors who do not want ORCID or wikidata.

6. Next steps

    a. Submit abstracts for the Biodiversity_Next conference.
    b. Expanding existing tests and pilots (look at options for including some of this work within SYNTHESYS+).
    c. BGBM Model: Participating institutes could select their most common collectors and add identifiers.
    d. MNHN Model: Participating institutes could try the protocol within their own collections for top collectors.
    e. Write up a list of the major challenges with example datasets for each type.
    f. Write a letter to CrossRef to recommend having the option of assigning an identifier other than ORCID for a person, ie Wikidata ID.
    g. Write a letter to ORCID to use identifiers other than GRID and ringgold for institutional affiliations, ie Wikidata ID
    h. Recommend that DiSSCo use ORCID as their sign-on for services.
    i. Review existing ORCID worktypes compared to what the community would like to recognise.
    j. Consider expanding use of ORCID sign on.
    k. Additional Wikidata records to be created in a systematic way such as entries found in, “Taxonomic Literature: A selective guide to botanical publications and collections with dates, commentaries and types (Stafleu et al.)”, https://www.sil.si.edu/DigitalCollections/tl-2/index.cfm - research and trial prior to Biodiversity_Next
    l. Assess current Wikidata entries and produce report on this for Biodiversity Next for people - consider a Short Term Scientific Mission (STSM) to achieve this.
    m. Adding person identifiers to the specimen data refinery in Synthesys+.
    n. Visualization of current information on identifier adoption and dissemination.
    o. What standards/changes do we need to accommodate person identifiers? For example, in Darwin Core.

ANNEXES:

A. List of Participants: Participants name, affiliation, expertise, contact

1. Elspeth Haston (RBGE) https://orcid.org/0000-0001-9144-2848 (e.haston@rbge.org.uk)
2. Arnald Marcer (CREAF) https://orcid.org/0000-0002-6532-7712 (arnald.marcer@uab.cat)
3. Quentin Groom (Meise Botanic Garden) https://orcid.org/0000-0002-0596-5376 (quentin.groom@plantentuinmeise.be)
4. David Shorthouse (CMN) https://orcid.org/0000-0001-7618-5230 (davidpshorthouse@gmail.com)
5. Nicole Kearney (BHL) https://orcid.org/0000-0003-2883-0906 (nkearney@museum.vic.gov.au)
6. Simon Chagnoux (MNHN) https://orcid.org/0000-0002-4210-484X (simon.chagnoux@mnhn.fr)
7. Vincent Boulet (BnF) (vincent.boulet@bnf.fr)
8. Chloé Besombes (MNHN) (chloe.besombes@mnhn.fr)
9. Nicky Nicolson (RBGKew) https://orcid.org/0000-0003-3700-4884 (n.nicolson@kew.org)
10. Josh Brown (ORCID) https://orcid.org/0000-0002-8689-4935 (j.brown@orcid.org)
11. Dominik Röpert (BGBM) (d.roepert@bgbm.org)
12. Rod Page (UoG) https://orcid.org/0000-0002-7101-9767 (Roderic.Page@glasgow.ac.uk)
13. Sarah Phillips (RBGKew) https://orcid.org/0000-0002-9155-8573 (Sarah.Phillips@kew.org)
14. Greg Riccardi (iDiGBio) (Greg.Riccardi@cci.fsu.edu)
15. Pavel Stoev (Pensoft) https://orcid.org/0000-0002-5702-5677 (projects@pensoft.net)
16. Teodor Georgiev (Pensoft) http://orcid.org/0000-0001-8558-6845 (preprint@pensoft.net)
17. Heimo Rainer (NHMW) https://orcid.org/0000-0002-5963-349X (heimo.rainer@nhm-wien.ac.at)

B. Extended Report

The drive to digitise the natural history collections of the world has the potential to link massive amounts of biodiversity and cultural data across time and place. However, the level to which this will enable the cross-discipline research and analysis that we aspire to is dependent on being able to make the connections effectively between the data. The long-standing issues of linking biodiversity data through people such as the collector, determiner, author have still not been fully resolved, although innovative approaches are being proposed and tested. There are options to manage authorities including the use of existing persistent identifiers for these entities and their utility and suitability need to be considered. Importantly, any system will need to enable links to contemporary and historic resources from all sciences and cultural disciplines.

This workshop brought together key contributors from across the relevant range of disciplines to agree a strategy for adoption and implementation for the authority management of people names in natural history collections and cross-discipline research and exposure.

Prior to the workshop, the participants gathered a set of fifteen use cases for the authority management of people names to help identify the requirements for any system in the future.

1. As a collections data manager, I want to unambiguously enter the primary and all other collectors as well the primary and all other determiners so that their efforts may be properly attributed.
2. As a collector of specimens, I want to create a virtual itinerary so that I can see where I have collected.
3. As a historian of science, I want to create a virtual itinerary of anyone's (alive or dead) collecting activities so that I can illustrate and publish on the impact collectors have had on locals' appreciation for nature.
4. As a collections data manager, I want to find duplicate specimen records across herbaria so that I can improve & fill the gaps in my institution's collections data with records pulled from elsewhere.
5. As a collections registrar, I want to unambiguously assess the expertise of an individual who has requested a loan so that I can trust them to handle specimens with an appropriate level of care and to return them in a timely fashion.
6. As an administrator at a national museum, I want to quantify the performance of my employees at managing the collections and their rate of progress in every quarter of the calendar year so that I can make informed decisions about staffing and salary.
7. As a collection curator, I want to consult online other handwritten labels of a collector to check my attribution is correct
8. As a museum employee, I want to report on the number of specimens and specimen data I handled the last quarter of the calendar year and compare that to the previous quarter.

9. As a past graduate student, I want to attribute specimens collected and identified by my mentors (some living, some dead) so I may honour their influence on my career path.
10. As a visiting taxonomist, I want my curatorial activities in foreign collections to be seamlessly and transparently documented (without me having to do any additional work) so that my institution receives recognition for having partially funded my travels.
11. As a historian I want to pull together the complete collections, correspondence, photographs, illustrations and publications of a botanist/zoologist/geologist so that I can write a biography of the individual.
12. As an exhibitions officer I want to be able to pull together all the related specimens and other works by an individual so that I can make a selection to present to the public.
13. As a lecturer I want to be able to access all the duplicates of a type specimen collected by a single individual to be able to teach students about nomenclature and taxonomy.
14. As a publisher of scholarly content and a register of DOIs for that content, I want to be able to add all present AND past authors to Crossref using persistent identifiers so that their contribution to biodiversity knowledge (their impact) can be tracked.
15. As a member of a research organisation, I want to be able to unambiguously link authors (present AND past) to our organisation so that the complete impact of my organisation can be tracked.

The workshop was structured into a series of three sessions, each of which comprised a set of short presentations and discussion.

Session One – Introductions and Use Cases (Collections)

A series of use cases from some of the major European collections were presented: Elspeth Haston (RBGE), Quentin Groom (BGM), Simon Chagnoux (MNHN), Dominik Röpert (BGBM) and Sarah Phillips (RBGK). These were to illustrate the current practices in how collections manage people names, the difficulties faced by collections and some of the work that collections are doing to incorporate people identifiers within their systems and to demonstrate the potential benefits.

It was clear from these presentations that collections are not following a standardised method of recording or presenting people names relating to specimens. The herbarium database at RBGE contains 14,000 individual names and substantial work was carried out to clean the data and enter additional information to help disambiguate similar names and to ensure that the correct collector is selected for the specimen being databased. However, when the information is presented on the online catalogue this disambiguation is not visible to the user, and they would therefore have to duplicate this work.

A similar situation was illustrated by a pilot carried out within the ICEDIG project where labels were being transcribed by a company in Suriname and assessed by RBGK and BGM.

This highlighted a key issue of handling collector teams, where the individual names were being either atomised or entered as a single string.

The discussion on these issues included the question of whether we should be entering names in a standardised format or whether we should be recording the name verbatim. The use of identifiers would enable the data to be captured verbatim but linked to a standard form.

Three pilot case studies from collections were presented. In BGM a trial was carried out to enter identifiers for a selection of the most common collectors in the collections database. These were initially linked to the identifiers in the Harvard University Herbarium (HUH) database. There were challenges in entering new names into the database and in downloading the data for use. It was therefore decided to test the use of Wikidata and VIAF. Over 900 names were processed within 18 hours, of which 175 names could not be successfully linked. The results were implemented as RDF on the BGM portal. Wikidata allowed rapid and easy fixing of duplicates and creation of new name records.

A demonstrator was developed by BGBM to show the potential of Linked Open Data. This used HUH, VIAF and Wikidata. The top 1,000 collectors were included of which 735 collectors were identified and these linked to about 50,000 specimens. Examples of two collectors, Humboldt and Brown, were presented, where the RDF were harvested from BGBM, BGM, RBGE, BHL, Wikidata and Europeana with impressive results.

In large institutes with diverse collections there are issues of fragmentation: different specimen collections, scientific publications, library, observations and national lists; and all these data held in many different information systems. At MNHN, most names are currently entered verbatim. A still ongoing pilot was carried out on to look at 500 of the most important names for MNHN across all the information systems. The discussion relating to the pilots also identified collector teams as a challenge. It was apparent that a significant impact on entering identifiers into the information systems could be achieved very efficiently. There was a significant long tail issue where an increasing amount of time would be required to process each name as less information was available for disamibiguation. It was agreed that machine learning could potentially play a large part in the process of disambiguating names.

Session Two – Use Cases (Literature)

From the literature community, there were representatives from several key organisations: Nicole Kearney (BHL), Vincent Boulet (BnF), Chloe Besombes (MNHN), Pavel Stoev (Pensoft) and Teodor Georgiev (Pensoft).

There is currently no authority control for names in the Biodiversity Heritage Library (BHL) but VIAF is being considered for mapping the names. It is estimated that about 40% of names in BHL could be associated with a VIAF identifier, and VIAF identifiers are now being added to Wikidata records automatically. Wikidata has been successfully trialled with the

model of creating pages with core metadata for the interested members of the public to adopt and enter additional data.

The Bibliotèque nationale de France (BnF) presented their current workflow which includes VIAF and ISNI. The BnF dataflow started with only manual cataloguing using the Archival Resource Key (ARK) protocol. Two years ago semi-automated cataloguing was added and the Atom Publishing Protocol (AtomPub) is now being used to send the data directly to the International Standard Name Identifier (ISNI) database. This is being used within a project to merge the two national French library name identifier systems. Dates, collaborative authors and publishers are being used to determine duplicates and cluster them. In case of doubt, merging does not occur and separate records are created. The International Standard Name Identifier (ISNI) is an ISO standard and is used more widely by a range of communities, including the music industry, rights management, etc. With regard to the General Data Protection Regulation (GDPR), there are exceptions for archival purposes with a cultural heritage perspective and ISNI is covered by this. There are also exceptions for public institutions holding a legal mission and BnF is covered by this. BnF is on the ISNI International Board, sharing a seat with the British Library (BL).

The above mentioned pilot study by MNHN to manage people name identifiers is linking internal databases by people names as a Data Proof of Concept project. A stable identifier (IDRef) is assigned to link to the French libraries database. This was discussed in the context of the ecosystem of the French academic libraries which are using the same systems.

Some of the challenges identified by the literature community included a reiteration of the problem of multiple names, whether in collector teams or multi-author papers. There was a question of whether the data needed to be clean before being published. The need to include more automated processes was emphasised with the development of scripts reducing the amount of human intervention. Registration to ISNI was discussed with several options being available for collections, including registering under the umbrella of a registration agency, e.g. British Library. Alternatively, the collection institute can become a registration agency itself. Registration agencies may be nationally or content scoped and can be a network of organisations. The concept of selecting the identifier based on its design purpose was generally agreed.

Session Three – Use Cases (Developers and Data Scientists)

The workshop included participants from the developing and data science communities who also discussed some of the current name authority management systems: Nicky Nicholson (RBGK, IPNI), David Shorthouse (CMN, Bloodhound, RDA, TDWG, GBIF), Rod Page (UoG), Josh Brown (ORCID), Greg Riccardi (iDigBio), Quentin Groom (Wikidata) and Vincent Boulet (VIAF, ISNI).

The International Plant Names Index includes the authors of plant and fungal names database which are in a standard form. Work to investigate how data can be mined from GBIF and linked by collector has shown the power of visualising the results. Analyses

included trends across different timescales, collection effort, collection clusters and duplicate discovery. The work to date has aimed to:

- Identify the problem: contrast the analyses possible when data are standardised (such as in the IPNI system) and the kinds of analysis that we would like to do at larger scale (GBIF mobilised specimen data) but currently cannot because the data are not comprehensively linked to a person standard.

- Use new technologies (data mining process based on unsupervised learning) to overcome what would otherwise be a very expensive data standardisation / linking process using minimal (read cheap) features from digitised species to assign these to a collector identity.

- Identify applications of this better linked dataset – as listed in the objectives above, quantify the number of post accession expert annotations that can be shared in a network of institutions sharing specimen duplicates, leading to shared curation. Higher level network analyses also possible, such as identifying sets of institutions which share many specimens and who could work together either to maximise overlap or complementarity.

The development of Bloodhound (https://bloodhound-tracker.net/) was aimed not only to enable biologists to receive professional credit for their work, but also to build on the emotional attachment people have to specimens. In order to enable recognition for taxonomists, for hosting institutions and for the taxonomist's home institution when they visit another collection, some key information needs to be recorded: IRI identifiedby, institutionCode, dateIdentified via TDWG standards and ORCID. However, some of these lines of relationships are not currently being recorded in a standard way. Bloodhound aims to make these connections by allowing people to self-identify and to make the connections for themselves and, potentially, for others.

The Biodiversity Knowledge Graph (BKG) needs globally unique persistent identifiers for the things we care about. There should also be easy discovery of these identifiers, structured data return, shared terms, the reuse of existing identifiers, and an attractive place to store and query structured data. Wikidata can be seen as a brokerage system to cross-link all the identifers from the various systems. By using Wikidata we would be building the resource where the community is already active rather than trying to entice the community into a new space. The aim would then be to build applications on top of Wikidata to provide the information that people need.

ORCIDs were designed to be identifiers for researchers, thereby creating a registry with a set of standard procedure for connecting reseachers to their affiliations and activities. It was created as an international-scale open research effort, and there are now more than six million IDs registered. The vision of ORCID is about people and their contributions and affiliations across time, disciplines and borders. The data in ORCID are owned by the individual and these control what happens to the data using authentication.

The Integrated Digital Biocollections initiative (iDigBio) is the national resource for digitised information about vouchered natural history collections. This is a network across the USA

for digitised data but also for the development of tools and online information resources. There are also plans to create an integrated network of networks which would see the collaboration of iDigBio, GBIF and ALA. People names have not been explicitly included in this proposal but there would be a clear advantage in terms of data quality checks and data cleaning.

A current challenge is the existing current resistance by people to use people name identifiers, including ORCID and Wikidata. Hence, there is a strong need to identify compelling use cases that have meaning for people and can encourage them to take-up identifiers.

Conclusions

No single system will be able to cover all the current and future uses for people name identifiers. Instead, we need to use the right system for the right purpose.

If we are using more than one system, then we need to ensure that we have a brokerage system in place. At present, Wikidata would seem to be the main brokerage system, although additional brokerage could be handled by VIAF and ISNI.

The priority systems for the community could therefore be:

- ORCID – self-created by living people. We need to encourage more take-up by our communities.
- Wikidata – all historic and living (with permission as required) authors, collectors, determiners etc, and an effective brokerage system. We need to build up the number of people records in Wikidata through core work and projects.
- ISNI – the ISO standard and an additional, formalised brokerage system with links to other communities. Identifiers for living people can be added without permission from the individuals.

As a community we need to make existing systems more robust, encourage adoption, increase coverage and build resources on them to increase motivation.

Major challenges

- Maintaining the momentum from the workshop will be important to achieve the progress planned for year two.
- Ensuring that the communication channels work effectively.
- The brokerage of multiple identifiers.
- Agreeing on a methodology for teams of people, e.g. collectors.
- Ensuring that the system can manage the order of names.
- The process of disambiguation for names.
- The needs and motivations of the authors, collectors and institutions.
- Ownership of your ID (for the living).
- Identifiers for living collectors who do not want ORCID or wikidata.

Actions coming out of the Workshop

- Submit abstracts for the Biodiversity_Next conference.
- Expanding existing tests and pilots (look at options for including some of this work within SYNTHESYS+)
- BGBM Model: Participating institutes could select their most common collectors and add identifiers.
- MNHN Model: Participating institutes could try the protocol within their own collections for top collectors.
- Write up a list of the major challenges with example datasets for each type.
- Write a letter to CrossRef to recommend having the option of assigning an identifier other than ORCID for a person, i.e. Wikidata ID.
- Suggest that DiSSCo use ORCID as their sign-on for services.
- Review existing ORCID worktypes compared to what the community would like to recognise.
- Consider expanding use of ORCID sign on.
- Additional wikidata records to be created in a systematic way -- research and trial prior to Biodiversity_Next.
- Assess current wikidata entries and produce a report on this for Biodiversity_Next -- consider a Short Term Scientific Mission (STSM) to achieve this.
- Adding person identifiers to the specimen data refinery in Synthesys+.
- Visualization of current information for adoption and dissemination.
- What changes on standards do we need to accommodate person identifiers?.

Links

Linnaeus in Wikidata: https://www.wikidata.org/wiki/Q1043

List of authors with Harvard Index of Botanists ID on Wikidata http://tinyurl.com/y4954yl5

JSTOR view of a collector's expedition: https://labstest.jstor.org/zambezi/

Demonstrator Harmonisation and semantic enrichment of collector names http://ww2.bgbm.org/herbarium/sparql.cfm

Picturae https://picturae.com/en/

BnF https://data.bnf.fr/en/12181133/patrick_blanc/

C. Other material

Workshop wiki: https://osf.io/qwegk/wiki/home/