The Naturalist's Friend

A case study and blueprint for pluralist data tools and infrastructure

Antranig Basman
Raising the Floor - International
amb26@ponder.org.uk

Abstract

Spreadsheets liberated individual end users to own their own data and curate its structure and relationships. However, most realistic individuals are embedded in multiple overlapping communities and societies, each of which brings different visions, ontologies and working practices. No modern digital tools perform the same liberating functions for communities that spreadsheets did for individuals. We will sketch the working patterns and relationships amongst some communities of naturalists that we have studied, distil these down into the description of a set of "challenge problems" and then sketch the design and infrastructure of a tool, codenamed "The Naturalist's Friend", This tool will enable multiple communities to collaborate on simple, tabular-structured data for which they have mismatched criteria for membership and interpretation of both rows and columns, without obliging any of them to compromise their autonomy over their own data standards, or lose the modern digital affordances of constantly live access to version-managed and publicly addressable incarnations of their shared data.

1. Introduction

Modern digital tools have an innate centralising, authoritarian tendency. Communities using them are obliged to converge on shared, central ontologies and workflows, or else be inflicted with inefficient and archaic data sharing practices such as emailing spreadsheets to one other. As communities grow larger and longer-lived, these archaic workflows eventually impose huge costs as it becomes impossible to determine what is the most recent, authoritative version of a resource for a community without costly and error-prone manual checking. The costs of filing and locating this data swamp all but the smallest communities operating on the simplest data.

1.1. Case Studies

We will present case studies of communities of naturalists cooperating to organise their data on observations and identifications of organisms, building on previous such studies in (Basman & Tchernavskij, 2018; Tchernavskij, Basman, Nouwens, & Beaudouin-Lafon, 2019). This area of data work is attractive because the absence of a direct profit motive enables economic concerns to be disentangled to some extent from commercial ones, and further because these communities already have centuries of experience in working with constantly shifting and constantly negotiated plural ontologies.

Following (Tchernavskij et al., 2019) we will cite larger-scale examples of extended naturalist communities trying to share more substantial sets of data (tens of individuals sharing thousands of species and observations) but also gain perspective by supplementing these with recent experiences in working on similar problems at a much smaller scale (two individuals sharing hundreds of records).

1.2. Challenge Problems

After tracing these sample communities, we will distil a collection of typical collaboration patterns into a selection of "challenge problems". In order to focus these problems and make the design space tractable, we will concentrate on the simplest practical data substrates, rectangular grids with a regular row structure, that space well-described by the ubiquitous .CSV files.

These substrates will prove quite adequately rich in intractable problems. Those which are more tractable relate to shared ownership over rows — communities may have different, and shifting standards for inclusion of records of relevance.

Less tractable are those which relate to the meaning and relations amongst columns. For example, almost every community of naturalists will include a column identifying the species of the observation. However, different communities will defer to different authorities that they recognise as competent to

1

assign such names, and even those which agree on a name may well disagree on the taxonomical interpretation of the name. Furthermore, the set of authorities relevant for a particular community will shift over time — in the face of such mismatches, all collaborating communities will want to stay in continuous contact, sharing access to the records of shared interest without loss of data or loss of meaning of data.

2. Small-Scale Case Study

In this section, we present a very small-scale example of data sharing, involving only 2 participants and 4 documents. The situation of collaboration focused on an upcoming visit by the two participants to Groton Wood in Suffolk, well-visited over 40 years by naturalist Oliver Rackham, whose visits had been carefully notated in 4 notebooks amongst the digitised archive held by Cambridge University Library[1]. The two participants were A, an experienced technologist with a weak amateur naturalist knowledge, and B, an experienced ecologist with good knowledge of the standard office tools (Excel, Word, etc.) that his profession involved. The purpose of the visit was to apply Rackham's observations to guide an appreciation of the wood's habitat, determine which of the species noted over the years by Rackham could still be discovered in the wood, and perhaps notate a few more.

2.1. Initial Compilation

Participant A devoted some time to deciphering the handwriting in the notebooks, and trying to compile an exhaustive index of each particular visit's sightings. This produced the first document, a very wide spreadsheet with 61 columns (as well as the observation data, one column for (nearly) each visit by Rackham to Groton) and 206 rows, one for each observed species. A selection of a later revision of this document, named "Document $A1_3$", is shown in Figure 1. In order to compile these data, A attempted to compensate for his lack of botanical knowledge by exploiting various online resources, especially the autocomplete facilities of websites such as BSBI[2] and iNaturalist[3].

As it transpired some weeks later, participant B had also compiled his own list, which was then emailed to A in the form of an Excel spreadsheet, a selection of which, named "Document $B1_1$", is shown in Figure 3. This spreadsheet has 3 columns and 146 rows. This immediately presented a somewhat interesting data normalisation challenge. Whilst the "scientific name" field should have been expected to be in common between the two documents, in practice there are quite some subtleties in this area. Whilst A had believed in adopting the BSBI's name (B's suggested online resource) for a taxon whenever it conflicted with iNaturalist's, they would end up with a result more reflective of accepted scientific research, it soon transpired that B's taxa were drawn from yet another source, (Stace & Thompson, 2019)[4]. Being the most recent version of the resource accepted by UK professionals, its contents were naturally accepted as authoritative for our application. However, given the further applications we desired for the data, it was necessary to maintain the linkage to the digital resources — this is the reason that A's larger spreadsheet shown in Figure 1 includes URLs for both of the online resources, and in a couple of cases the three resources referenced for a species each refer to it by a different name.

---

[1] available at `https://cudl.lib.cam.ac.uk/collections/rackham/1`

[2] The Botanical Society of Britain & Ireland, whose Online Atlas of the British and Irish Flora is available at `https://www.brc.ac.uk/plantatlas/` — this was already a resource suggested by participant B

[3] Primarily a citizen science platform, already described in (Basman & Tchernavskij, 2018; Tchernavskij et al., 2019), whose taxon interface is available at `https://www.inaturalist.org/taxa/`

[4] This a bulky and expensive paper reference, very recently published, whose information at the time of writing is still not widely available in digital form, and which is now even temporarily out of print due to its extreme popularity.

Figure 1

Groton Wood Species List from Rackham's Notebooks — File Edit View Insert Format Data Tools Add-ons Help — All changes saved in Drive

Ordinal	Rackham's notation	Species/Genus	Common name	iNaturalist Link	BRC link	AWI	Obs	DV	11 Apr 2010	3 Sep 2010	20 Apr 2011	5 Aug 2011	3 Aug 2012	17 Apr 2014	25 Jul 2014
1	[Lapsum?]	Lapsana communis	Nipplewort	https://www.inaturalist.org/taxa/55981-Lapsana-communis	https://www.brc.ac.uk/plantatlas/		8	1							1
2	Cirs vulg	Cirsium vulgare	Common thistle	https://www.inaturalist.org/taxa/52989-Cirsium-vulgare	https://www.brc.ac.uk/plantatlas/		11	1						1	1
3	Chamaener	Chamaenerion angustifolium	Rosebay Willowherb	https://www.inaturalist.org/taxa/564969-Chamaenerion-an	https://www.brc.ac.uk/plantatlas/		20	1		1					
4	Scroph nod	Scrophularia nodosa	Common Figwort	https://www.inaturalist.org/taxa/124485-Scrophularia-nodo	https://www.brc.ac.uk/plant	0.5	12	1	1	1	1				
5	[Geum viv?]	Geum rivale	Water Avens	https://www.inaturalist.org/taxa/55750-Geum-rivale	https://www.brc.ac.uk/plant		1	1	1				1		
6	Ranun rep	Ranunculus repens	Creeping Buttercup	https://www.inaturalist.org/taxa/48229-Ranunculus-repens	https://www.brc.ac.uk/plantatlas/		5	1							
7	Dryop cf affinum	Dryopteris affinis	Scaly Male Fern	https://www.inaturalist.org/taxa/321288-Dryopteris-affinis	https://www.brc.ac.uk/plantatlas/		1	1							
8	Arctium	Arctium (lappa/minus)	(Greater/Lesser) Burdock	https://www.inaturalist.org/taxa/75501-Arctium-lappa	https://www.brc.ac.uk/plantatlas/		15	1 - as		1				1	
9	[Puten ster?]	Potentilla sterilis	Barren Strawberry	https://www.inaturalist.org/taxa/68853-Potentilla-sterilis	https://www.brc.ac.uk/plant	1	9	1						1	
10	Senec eruc	Jacobaea erucifolius	Hoary Ragwort	https://www.inaturalist.org/taxa/168817-Senecio-erucifolius	https://www.brc.ac.uk/plantatlas/		9	1		1	1				1
11	Dactylis	Dactylis glomerata	Cock's-Foot	https://www.inaturalist.org/taxa/52720-Dactylis-glomerata	https://www.brc.ac.uk/plantatlas/		4	1	1						
12	[Gulimerc?]						1								
13	Hyp pfor	Hypericum perforatum	Perforate St. John's Wort	https://www.inaturalist.org/taxa/56077-Hypericum-perforat	https://www.brc.ac.uk/plantatlas/		5	0							

Figure 1 – Participant A's data collection (in final form) - Document A1₃

Figure 2

Visual checklist of AWI from Rackham's Groton Wood notebooks — File Edit View Insert Format Data Tools Add-ons Help — All changes saved in Drive

Ordinal	Rackham's notation	Species/Genus	Common name	AWI	Obs	G	H
5	[Geum viv?]	Geum rivale	Water Avens	1	1		
79	Ranunc auricomus	Ranunculus auricomus	Goldilocks Buttercup	1	1		

Figure 2 – Participant A's "illustrated guide of highlights for amateurs" - Document A2

Figure 3 – Participant B's original data collection - Document $B1_1$

Figure 4 – A's importation of B's $B1_1$ as $AB1$ for harmonisation (final version $AB1_3$)

2.2. Rounds of Harmonisation

A did the work to normalise B's authoritative names into column B of his spreadsheet which then made possible the corresponding normalisation of determining which species he and B actually had in common — in practice, both transcribers had missed different sets of Rackham's observations (which also in themselves sometimes diverged in species name from all the authorities in play). This normalisation had to be done by time-honoured, manual techniques — A's spreadsheet $A1$ was enhanced with a boolean column "in DV" and B's spreadsheet B_1 was imported by A as $AB1$ (Figure 4) with a new column "in AB", each was sorted into alphabetical order and the checkboxes in the opposing spreadsheet were checked off. At this point, further discrepancies between A and B's ontologies for the problem surfaced. In A's transcription, as a result of his domain inexperience, two large categories of species had been systematically omitted — i) all trees, ii) all extremely common species with which he was familiar, e.g. nettles, brambles and the like. This was on the basis that such species would be both so frequently mentioned in the notes, and frequently encountered in the field, that observing them would have little informational value. On the other hand, B's professional experience had trained him that no such observation should be considered beneath notice. However, on the other side, A had also transcribed all fungi from Rackham's notebooks, which B had omitted since i) this was an area outside his professional expertise, and ii) since the proposed visit was in spring, few of these would be visible. On the issue of the trees, A decided for the while to persist in his stubbornness and constructed an extra annotation T in his working copy of B's spreadsheet in order to encode that a species was missing in his own for this reason. A then projected his excess rows into B's schema and shared with B via Google Sheets an enhanced version of their original spreadsheet in their schema, with the additional observations filled in. In practice this led to two elaborated resources, $A1_2$ and $AB1_2$, whose columns were supersets of those present in the original $A1_1$ and $AB1_2$. A then went through a further round of transcription from the

4

original notebooks in order to turn up not only all of B's excess observations, but a couple further that had been missed by both transcribers, resulting in $A1_3$ and $AB1_3$.

2.3. Compilation of Derived Resources
In this section we now describe two further resources that were derived from the basic transcriptions $A1$ and $B1$.

2.3.1. Visual Checklist for Amateurs
At this point it was time to use the pooled information to derive two further kinds of artefact, primarily to be used in the field at the actual visit. The first kind was an illustrated guide to a couple of dozen of the most frequently observed "Ancient Woodland Indicators" (AWIs)[5], since as well as A, three and a half other amateur naturalists would be accompanying on the trip, to whom the species would be largely new. A selection of the available images for the taxon available on iNaturalist via column E of resource $A1$ was examined, and A selected two which in his opinion gave an impression both of the overall habit of the plant and then some detail which he found interesting. This gave rise to a derived resource $A2$ (shown in Figure 2 by A from document $A1$, which corresponded to only a small subset of the rows from $A1$ - in this case, to those which had been marked as AWIs, and fell into the top 25 of species when sorted by number of mentions in Rackham's notebooks.

This process of image selection is an interesting one to which we will return in the next section, but it is well worth noting that the level of experience of the transcriber will have a significant bearing on the choice and value of the resulting selection. An inexperienced transcriber may well end up selecting an image which is visually appealing, but is useless for resolving some of the significant details which would have to be used to make an reliable identification of the species in the field. On the other hand, an expert botanist might well end up selecting images which were less visually appealing but showed significant details, but on a yet further hand, an expert botanist may well not compile such a visual list at all since experts tend to derive their identifications from written descriptions and professionally compiled dichotomous keys, and this checklist $A2_1$ itself was indeed primarily aimed at the needs of amateurs. However, this highlights that such a visual checklist is a form of "community resource" and a particular selection may well end up serving the needs of some communities of interest better than others.

2.3.2. Full Checklist for Observations
A and B on their actual visit to Groton would need an uncluttered template document in which to enter their observations in the field. This would include the bare minimum information from the transcription documents $A1$, $B1$ of the (or a) scientific name for the species, its common name, AWI status, and then room in which to write notes of the observation. A drew up his own version of this resource, named $A3$ which is shown in Figure 5.

B in the meantime drew up their own rendition of the checklist for his own use, which is based on essentially the same data but with fewer columns, named $B3$ shown in Figure 6. It's worth noting that whilst by this point the discrepancy between the two checklists with respect to inclusion of trees had been resolved, that B once got to the field expressed disappointment that somehow the common species that had been in their original collection $B1_1$ had gone astray from their $B3$. This illustrates the huge potential for accidental loss or corruption of data as it passes through clunky, manual workflows such as these since each instance of the data is only as good as the last visual check made to correlate it with its informationally disconnected sources.

2.4. Information Flow Between Documents
The complete information flow network between the different documents in this collaboration is illustrated in figure 7. It would clutter the diagram unduly to try to visually represent the detailed nature of the data relationships, but the previous discussion should make clear that as well as straightforward cases where extra columns or rows are introduced into upstream resources, there are several instances of

---

[5]Plants whose presence which, on account of their extremely slow colonisation of adjacent forest, marked the area they were found in as "ancient woodland", that is, likely to have been continuously occupied by woodland for upwards of 400 years.

Figure 5 – Participant A's checklist for the field - Resource A3

*Figure 6 – Participant B's checklist for the field - Resource B3*

complex multilateral relationships. An example of such a relationship is between $AB1_3$ and its parents $A1_3$ and $AB1_2$ (and ultimately $B1_1$), where neither the rows nor columns of these documents are proper subsets of either of their parents — along each axis, some material has been "inherited" from one parent, then possibly some of that information has been selected away some policy, and then enlarged by a subset taken from another parent. It is possible that a more or less regular description of this policy could be expressed in some declarative form, perhaps drawn from the domain of "lenses" promoted by the bidirectional programming framework of (Foster & Pierce, 2009). However, we should, in the light of this particular process, note some further subtleties.

Even in our sample community of 2 participants, we find that the issues of authority and relevance are central. As the domain expert, any of B's data decisions should expect to take priority over any made by A, and one would expect the natural structure of data updates within this network to reflect this. In fact, that B's updates took the form of emailed spreadsheets ensured that no data that he "owned" might get modified by A, except inadvertently. One would expect that any technological solution to such collaboration problems would be able to make it inherently clear to all participants what the nature of such authority relations might be — for example, in this situation both A and B would both be mutually, implicitly clear that A would perform no actions that might cause data notionally "owned" by B to be updated, unless B had explicitly requested or confirmed this.

One suspects that the difficulty in reflecting such complex, and possibly shifting relations of ownership and authority over different subsets of the data is what drives most collaborators back to the "clunk workflow" of emailing spreadsheets, since at least in that model it is clear what data has been transferred and no possibility of it shifting under the participants' feet. However, even in this small-scale collaboration the lack of technological support led to considerable inefficiencies and numerous errors, many of which have probably not even yet been characterised.

This collaboration is an endless story of "broken relations" — each of the 5 documents in their various incarnations morally is connected to the others via what should be relations, and should any of them become updated, one would hope it would be possible to choose through some automated means to propagate the corresponding updates to the others. For example, should a fresh decision be made about the accepted species name for an element of $B1$, that this could be propagated by some automated process to $A1$, $A2$ $A3$ and $B3$. Right now this update would have to be copied by hand between the involved documents.

However, it would be obviously inappropriate for these updates always to happen transparently and automatically, as it would if Figure 7 represented a conventional kind of "dataflow network". Each author has their own interest in the integrity and meaning of their documents, and whilst in some cases

6

Figure 7 – Complete information flow between documents in small-scale collaboration
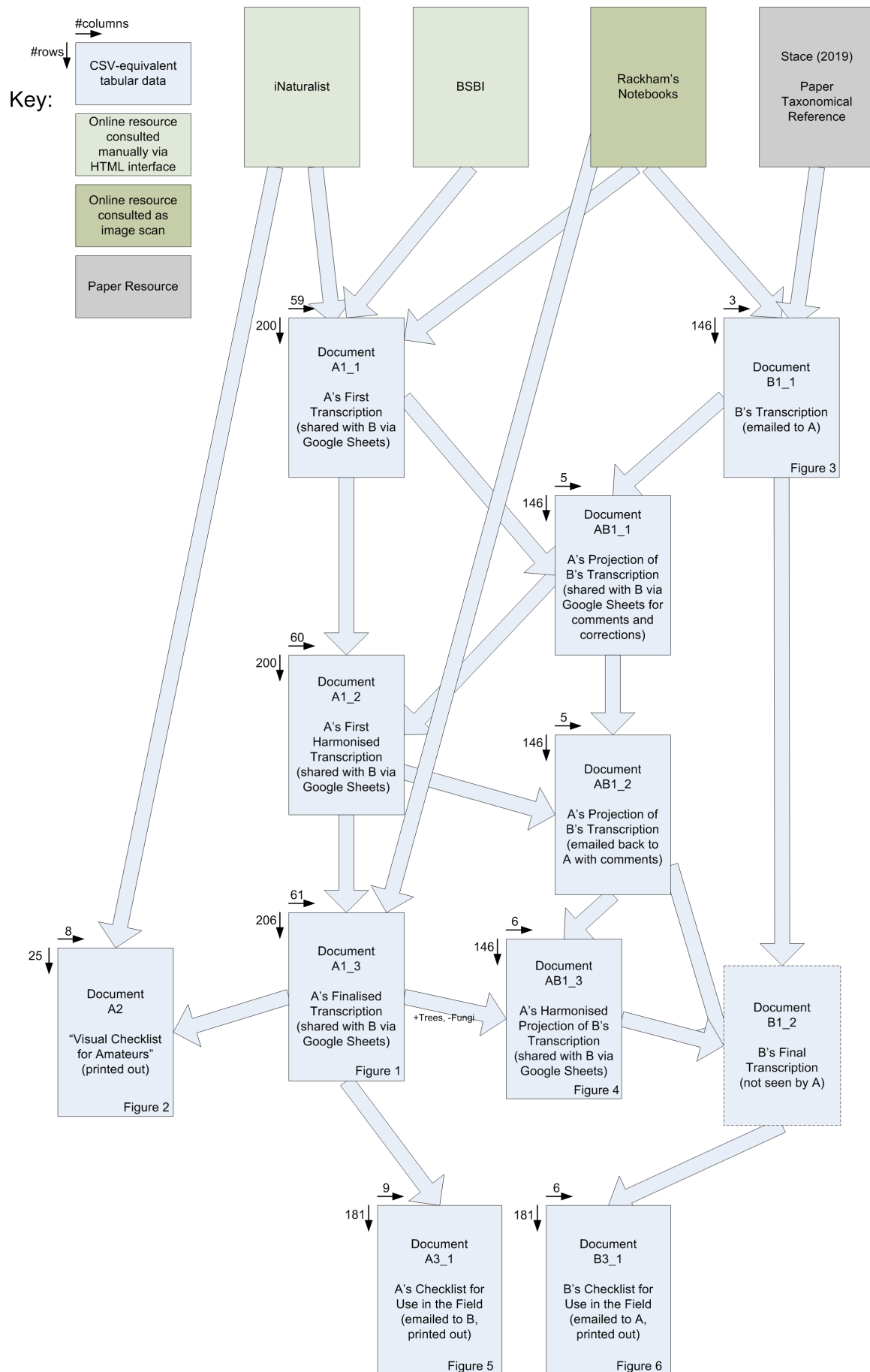
they might want to set a policy of automatically accepting certain kinds of updates, they would want to preview them and perhaps temporarily or permanently reject them.

3. Larger-Scale Case Study

In this section we will consider the case of a larger group of collaborators, a increasingly connected group of naturalists working in the Salish Sea, a network of waterways stretching between British Columbia and Washington State and encompassing Vancouver Island and several neighbouring islands. Primarily these groups are composed of professional naturalists, collecting much larger datasets than those in section 2, typically of thousands or tens of thousands of observations over a timescale of years or decades. Participants from these groups consult a variety of specialised online and paper resources to derive taxonomies and checklists of species, and collect observations through a variety of methodologies recorded in a variety of formats. Some of the goals of the growing collaboration are to produce tools where some core of the community data may be rendered into some common forms in order to facilitate statistical and visual studies of species occurrence across the region as a whole, without forcing the communities to compromise on local imperatives for data collection and representation. A representation of some of the communities involved and some data that they have shared, together with some upstream information sources is shown in Figure 8.

Given the size and number of the communities we are representing here, Figure 8 suppresses numerous details that would be visible had we tried to draw a diagram of detailed interactions in these communities at the same scale we have shown in Figure 7. In conversations with the members of these communities, both those pursued for (Tchernavskij et al., 2019) and privately, we have seen that on each occasion these communities interact over their shared data, tangles such as that shown in 7 and much worse will routinely arise.

3.1. An ecology of emailed spreadsheets

The working practices shown in 2 are found to be typical in this much larger, more long-lived community, only the costs imposed are more severe. Spreadsheets with thousands of entries are routinely mailed back and forth between participants, all with divergent schemas and information contents, and referred to discrepant taxonomies. The task of normalising one of these documents with respect to the information standards of a different community may be a a long-boiling task of several weeks rather than just an evening annoyance. One of the participants has explained that their community is sitting on a file server holding decades worth of such documents whose contents may never again be understood or indexed unless they are lucky enough to experience the transient tenure of a graduate student with the relevant skills who might succeed in mining a handful of them. Since we've given the flavour of such problems in section 2 we will draw a veil over the details except to talk a little further of an domain issue which is exposed more sharply in this larger-scaled community, that of taxonomies.

3.2. Taxonomies in the large

Given these are peer communities of experts, the issue of taxonomical choices becomes more substantial than it was during our small-scale collaboration, and exposes some of the core requirements on our "Naturalist's Friend" application that we will try to sketch out in section 4, that go beyond those of data tools so far developed. Figure 8 shows some of the many resources that our communities might consult for taxonomical information, and not shown are many further resources that might exist only on paper, or even in the form of personal contact with individual subject-matter experts who might answer questions from their own personal experience. Whilst there is obvious value in centralising as much shared knowledge as possible in a centralised "taxon resolution database" (TRD) as shown in the centre of the Figure, it's clear that, as with, for example, iNaturalist's taxon database itself, no single choice is going to be wholly valid for all participants. The level of databases highlighted with question marks in the row below the central TRD suggests that each community member would in theory expect to administer choices about a TRD local just to their project. A particularly salient example of this is the Metchosin project, which, given it collaborates particularly often with federal authorities publishing lists of endangered species drawn from a nationally standardised taxonomy, needs to be able to refer records
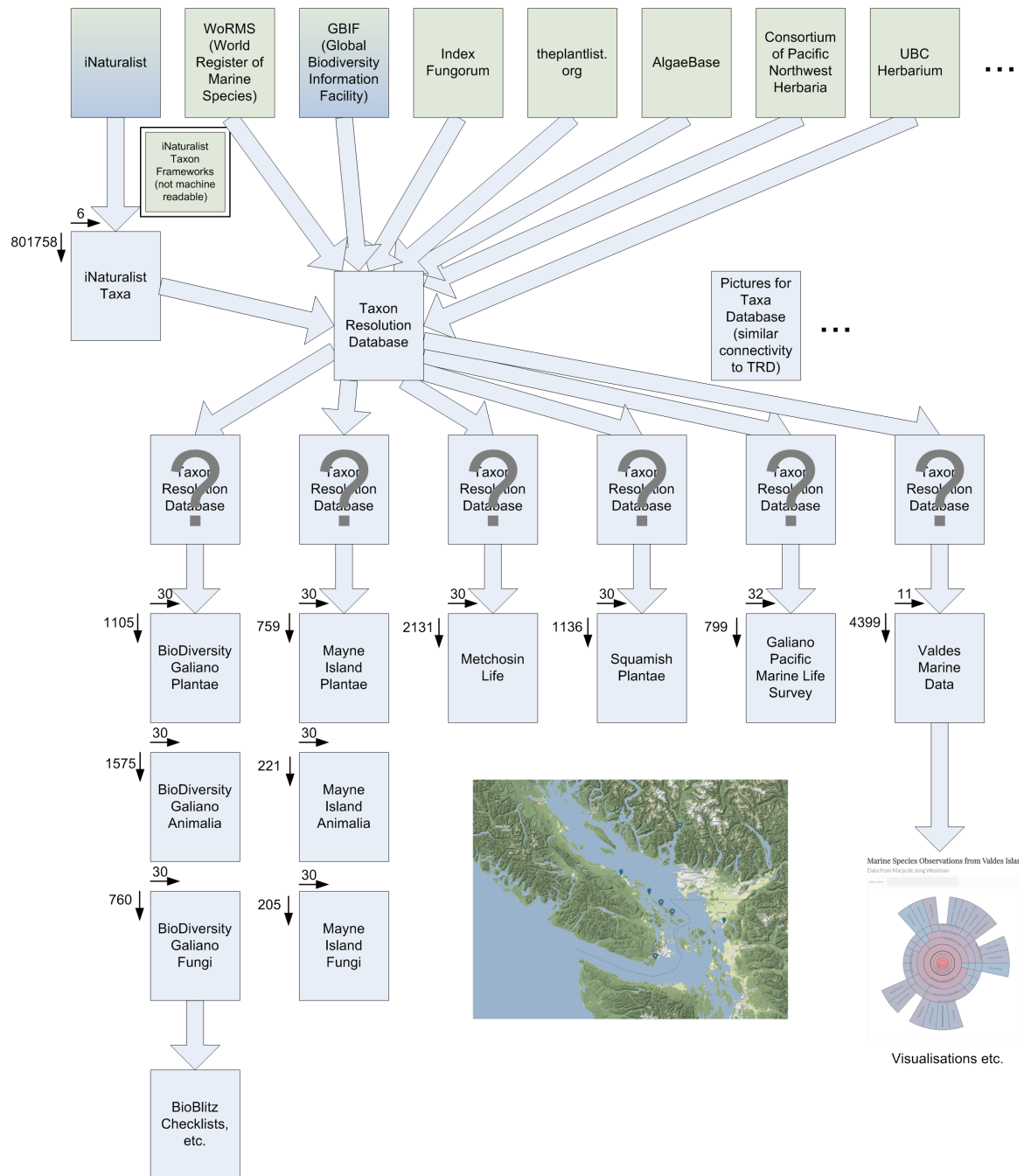
Figure 8 – Partial information flow between documents in large-scale Salish Sea collaboration

to that taxonomy. But in practice, it would be impossible for any participant to completely administer a local taxonomy and express design choices, for example, about a list of 800,000 species — instead, they will want to issue strategic directives at a high level of granularity, e.g. "For all organisms within this taxon we defer to authority A's published list, and within this other taxon we defer to authority B with the exception of the following 5 species for which we override their choice".

This is in fact a problem which the iNaturalist platform itself contains facilities for managing, in its "Taxon Frameworks" system[6]. Unfortunately this system has a few drawbacks. Firstly, it is implemented only with respect to iNaturalist's taxonomy itself, and its encoding of the taxonomic relationships is not made public in a machine-readable form to external users (represented in Figure 8 by the black box drawn around its document). The Taxon Frameworks system can't be used by 3rd parties to administer taxonomies for their own communities, without forking and hosting an entire iNaturalist instance devoted to their community. This is prohibitive for reasons discussed in (Basman & Tchernavskij, 2018; Tchernavskij et al., 2019). Secondly, the Taxon Frameworks UI is rather intricate, a little clunky since lightly used by most iNaturalist users, and aimed towards experts who are thinly represented in their userbase. Nonetheless, the functionality it offers is a useful model for affordances we would want to democratise. Useful HTML renderings of its taxonomic mappings and their relationships to upstream authorities are available at URLs such as `https://www.inaturalist.org/taxa/1/taxonomy_details`.

4. Design Sketch

In this section we will try to sketch out the parameters of our proposed application and data infrastructure, codenamed "The Naturalist's Friend", aimed at mitigating the kinds of problems we've described in sections 2 and 3.

The interface of our desired tool will physically resemble as closely as possible that of one of the widely deployed spreadsheet tools (Office, Numbers, Google Sheets, etc.) in order to allow users to leverage the skills and experience they've gained in these end user tools. We will simplify our initial design to a great extent by removing the facility for any formulae or traditional "computed values" since these are not of particular relevance to our communities of interest. However, instead, we will augment the interface as a whole by extra controls which allow the community to make and break relations to other similar data — either data about related but distinct items, such as the taxonomy of observations, or data about the same class of item, such as observations made by other communities. These relations will always be "forgiving" in that violations of the relation will not be treated as an error condition by the interface, but just as a further feature of the data environment which the user may choose to highlight. Since in addition there will be ready access to high-integrity snapshots of past versions of the data as well as its accompanying relations, the effect of accepting or rejecting potentially destructive bulk updates of the data induced by data can always be undone as easily as they are made.

However, following (Basman & Tchernavskij, 2018), it will be impossible to specify the interface of our tool in separation from its infrastructure and externalised affordances — since it is a crucial part of its specification that it must work for its communities "in place". For example, at any time a user should have instant access to the basic substrate of their data — in this case, a .CSV file at an easily determinable URL or direct download. Further, there should be similarly directly externalised access to the version history of the community's own data in a publicly intelligible format — we recommend a standard `git` repository, and furthermore to mirror community accounts and relations by corresponding ones within `github`'[7]. In addition, the packaging and distribution of the data itself will also be geared through a widespread repository system such as `npm`, where the data is augmented with a simply-structured JSON metadata block in a format perhaps derived from that of the Frictionless Data project from Open Knowledge Labs [8].

---

[6]Described at `https://www.inaturalist.org/pages/taxon_frameworks`

[7]This is not an irreversible design choice — communities should be able to host their own storage and versioning apparatus, but a useful default would assign them to these, the most widely deployed technology choices for this class of problem

[8]`http://okfnlabs.org/projects/frictionless-data/`

A crucial community-oriented principle of design is that the tool must pay its way even in a community where the individual user is its only adopter. This implies that it must behave no worse than a traditional spreadsheet in any aspect that is relevant for its use in the planned communities, and hopefully it will behave better in many aspects. This implies that the traditional archaic workflow where an incoming collaboration takes the form of a spreadsheet emailed to the user must be a centrally designed use case. Another implication of this principle is the approach to "forgiving relations" mentioned in an earlier paragraph — the presence of a relational constraint in the system should not make it less usable for the purposes of plain, literal data entry than it was before.

With this, we end up with an interface that resembles, in one pane, a highly simplified spreadsheet application, which is only capable of working one CSV-equivalent literal data, but which presents another pane, the "community pane" showing a view structurally similar to that in Figure 7, showing the relationships of the currently focused data to related documents in the community, some of which will be documents in different schemas managed by this or other users, and some of which will be significant different versions of the current document. The notion of "significance" will typically be that two community members have interacted by exchanging or signing off the document version for some purpose.

4.1. "Make Relation" and "Pull Relation" actions

The interface will support some interesting operations not typically seen in office applications. The most important of these is the "make relation" operation. This will, given another resource either stored on disk or available on the network, instruct the system that it bears some relation to the current document. This relation will usually take the form of selecting some subsets of columns in the corresponding documents, and selecting from a set of available "lenses" which map the values one onto the other. An example of such a lens might be a "taxonomical lens" which maps the names of species used in one community to those used in another, with reference to one of the taxon resolution databases described in section 3.2. Another example might be to map geographical coordinates expressed in decimal latitude and longitude as used by modern cloud-based maps such a Google Maps into UK Ordnance Survey coordinates which have been used by historical and professional communities.

Along with defining such lenses for values, the interface will allow the user to select or define a mapping which explains the presence or absence of rows in the document based on functions of the column values — e.g. "The version of this data held by community B does not include any of our rows for which the value of column `kingdom` is `Fungi`. An instance of such a relation is the one connecting documents $A1_3$ and $B1_3$ seen in section 2. A more ambitious example would be to ask the system to synthesize a first pass at document AB_1 given documents $A1_1$ and $B1_1$ — in this case, the schemas of the parent documents are quite different, and the taxonomies are also misaligned.

It's an overriding design principle that these relations/lenses will not be hard relational constraints such as those seen in relational databases, or more recently in the "linked data" initiative described at `http://linkeddata.org/`, but soft constraints that the only consequence of violating will be that the data not respecting the constraint will be highlighted in some views — the "forgiving relations" described in the opening of this section. However, whenever an author is happy with the consequences, or merely as an experiment, they may choose the "Pull Relation" action for such a lens, which will update their own document with the corresponding mapped values. Given every document in the system is durably version managed, these experiments can be reversed as quickly as they are made. An example of this might be for the owner of document $A2$ to pull data from, say, an updated $A1_4$ into their schema. This might cause a species for which they had previously selected an image to fall out of the rankings, and be replaced for another, a piece of data loss they would have to patch up manually.

4.2. "Impute Version" action

A variety of the "Make Relation" action, that we call "Impute Version", can occur when a document is imported into the system from the outside, e.g. when received as an emailed spreadsheet, and the document is to be considered an updated or otherwise variant version of one already in the system. Normally, all the "lenses" involved in this imputation will be the null lens, but if the data involved is not

well supplied with unique keys or there has been some other kind of disruption to its structure, the work of mapping the rows of the updated data onto the original may not be trivial. Just as with the ordinary "Make Relation", we expect the system to apply brilliant AI, or else draw on a repertoire of canned responses previously found useful in the community, to ease importing the data back into alignment. An example of such an action would be the arrival of document AB_2 in the section 2 system — this is reasonably well-aligned with earlier versions but in practice required significant manual effort to check and align.

4.3. "Break Relation"
The correlate of the "make relation" action is required when a community's direction seems to take it away from alignment with one of its previous neighbours. Our small-scale example doesn't span enough time or space to encounter this, but one could imagine this dynamic emerging, for example, with the taxon resolution databases in section 3. Two communities may have been content to share a database for a significant time, until at some point either a change of mission, or the discovery of data anomaly imposes on them the burden to start maintaining separate databases. At this point it should be easy both to simply "fork" a pre-existing resource into two which may then continue to diverge, but also to retroactively "unpull" any data which had previously been pulled over a previously existing relation to restore it to the value it would have otherwise had, without disturbing any other data in the system.

4.4. System Infrastructure
We've spoken loosely of "the system" in previous sections but, following the initial discussion in this section, it's clear that it could not be incarnated as either a purely localised or a purely distributed system. Rather than being a "walled garden" system as is operated in typical data notebook systems, data is not only imported into the system from the net at large, but is also automatically exported back into it at easily discoverable URLs. For every document at every point in the system, the interface allows easy to a public URL from which an ordinary CSV for that document's state can be read. However, alongside this export must similarly be an encoding of the metadata not only for any schemas in the data, but also of the system's tracking of relations to related data. This will be done "by convention" — given a particular URL scheme, the URL for the metadata can be statically deduced from the URL for the CSV. If the metadata is not found at that URL, then the data will be imported by copying into the user's "local node" of the system, with an annotation in its metadata encoding where the data was found in the wild.

5. Conclusion
We've walked through a number of real-world collaboration problems, and exhibited numerous use cases that are poorly met by existing office-type tools, and all others that are currently available to our target communities at a price they can afford. Our design sketch for a family of solutions needs considerable refinement down to detailed user interactions and detailed implementation specifications, but we believe represents the blueprint of a relatively tractable implementation effort, and one that could ease the huge burden of articulation work still felt by cooperating communities sharing related but discrepant data. We've stressed the importance of operating a rarely seen idiom, which we've named "forgiving relations", under which the system indefinitely tolerates data which does not satisfy relationships which have been encoded into the system.

6. References
Basman, A., & Tchernavskij, P. (2018). What Lies in the Path of the Revolution. In *Proceedings of the Psychology of Programming Interest Group*.

Foster, J. N., & Pierce, B. C. (2009). *Boomerang Programmer's Manual*. Retrieved from `http://www.seas.upenn.edu/~harmony/manual.pdf`

Stace, C., & Thompson, H. (2019). *New Flora of the British Isles* (4th ed.).

Tchernavskij, P., Basman, A., Nouwens, M., & Beaudouin-Lafon, M. (2019). Control and Ownership of Artifact Ecologies in a Biodiversity Research Network. In *To appear: Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work*.