SYNTHESYS+ Virtual Access - Report on the Ideas Call (October to November 2019)

Helen Hardy, Sandra Knapp, E. Louise Allan, Frederik Berger, Katherine Dixey, Bernadette Döme, P.-Y. Gagnier, Jiri Frank, Elspeth Haston, Joachim Holstein, et al.

OPEN ACCESS

CrossMark

Project Report

SYNTHESYS+ Virtual Access - Report on the Ideas Call (October to November 2019)

Helen Hardy‡, Sandra Knapp‡, E. Louise Allan‡, Frederik Berger§, Katherine Dixey|, Bernadette Döme¶, Pierre-Yves Gagnier#, Jiri Frank¤, Elspeth Margaret Haston«, Joachim Holstein», Steffen Kiel^, Maria Marschler˅, Patricia Mergen¦,ˀ, Sarah Phillips ͻ, Rivka Rabinovich₵, Begoña Sanchez Chillónℓ, Martin V Sorensen ᗋ, Marco Thines P, Maarten Trekels¦, Robert Vogtₐ, Scott Wilson‡, Karin Wiltschke-Schrotta˅

‡ Natural History Museum, London, United Kingdom
§ Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
| The Alan Turing Institute, London, United Kingdom
¶ Hungarian Natural History Museum, Budapest, Hungary
# Muséum National d'Histoire Naturelle, Paris, France
¤ Národní Muzeum (National Museum), Prague, Czech Republic
« Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
» State Museum of Natural History Stuttgart, Stuttgart, Germany
^ Swedish Museum of Natural History, Stockholm, Sweden
˅ Natural History Museum Vienna, Vienna, Austria
¦ Meise Botanic Garden, Meise, Belgium
ˀ Royal Museum for Central Africa, Tervuren, Belgium
ͻ Royal Botanic Gardens Kew, Surrey, United Kingdom
₵ The Hebrew University of Jerusalem, Jerusalem, Israel
ℓ Museo Nacional de Ciencias Naturales-CSIC, Madrid, Spain
ᗋ Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
P Biodiversity and Climate Reserch Centre (BiK-F), Frankfurt, Germany
ₐ Botanical Garden and Botanical Museum, Berlin, Germany

Abstract

The SYNTHESYS consortium has been operational since 2004, and has facilitated physical access by individual researchers to European natural history collections through its Transnational Access programme (TA). For the first time, SYNTHESYS+ will be offering

virtual access to collections through digitisation, with two calls for the programme, the first in 2020 and the second in 2021. The Virtual Access (VA) programme is not a direct digital parallel of Transnational Access - proposals for collections digitisation will be prioritised and carried out based on community demand, and data must be made openly available immediately. A key feature of Virtual Access is that, unlike TA, it does not select the researchers to whom access is provided. Because Virtual Access in this way is new to the community and to the collections-holding institutions, the SYNTHESYS+ consortium invited ideas through an Ideas Call, that opened on 7th October 2019 and closed on 22nd November 2019, in order to assess interest and to trial procedures. This report is intended to provide feedback to those who participated in the Ideas Call and to help all applicants to the first SYNTHESYS+Virtual Access Call that will be launched on 20th of February 2020.

Keywords

access; collaboration; digital data; natural history collections; virtual data; digitisation; digitization;

Introduction

In the past decade, great changes and advances in digital, genomic and information technologies have taken place, supporting new paradigms of research on natural science collections. SYNTHESYS has been a critical instrument supporting this transformation. Since 2004, SYNTHESYS has underpinned new ways to access and exploit collections, providing critical new insights for thousands of researchers, while fostering the development of new approaches to face urgent societal challenges. SYNTHESYS+ (Smith et al. 2019) acts as a fourth iteration of this programme as it evolves into a sustainable and independent Research Infrastructure through the DiSSCo (Distributed System of Scientific Collections) ESFRI initiative (Addink et al. 2019; Lannom et al. 2019). The SYNTHESYS+ programme brings collections institutions together with the European branches of the global natural science organisations to address the emerging challenge.

For the first time, building on preparatory research in SYNTHESYS3, SYNTHESYS+ will create and provide an entirely new service class that allows access to virtual collections, freely available to a global user community, through its Virtual Access (VA) programme. Two calls for VA will run, the first in 2020 and the second in 2021; applications will be prioritised by an independent panel. This VA programme will then fulfil prioritised requests, by funding relevant digitising (collections-holding) institutions to undertake the necessary digitisation, data curation and provisioning to make the data accessible through open public portals.

Digitisation in the SYNTHESYS+ VA programme is defined as a request for collections data in digital form. This may be digital images, digital data (including that from 3D scans and other complex technologies), or digital molecular or chemical data. Requesting molecular or chemical data may significantly increase the costs of any proposal, and

significantly limit the institutions able to act as digitising partners at this stage, but the programme allows for more innovative pilot proposals where the digitising institutions are able to support these.

It is intended that data generated as part of the SYNTHESYS+ VA programme will be served through the existing network of public portals (principally GBIF and institutional data portals) to provide unfettered access to collections data and associated media. These portals already have a very extensive user base and are independently maintained through core activities of the partner network. The primary constraint on the exploitation of these portals to help address societal challenges is the data they contain, as this usually represents a small fraction (4.5-18% for the eight largest collections in the consortium) of any given institution's physical holdings. VA within SYNTHESYS+ is designed to help address this critical gap, increasing the proportion of collections freely available digitally, and targetting this in response to evidence of a strong demand by current, emerging and new user communities.

Nineteen of the SYNTHESYS+ partners are participating in the VA programme. As part of the VA programme, digital data from a wide variety of workflows can be requested: standard 2D imaging (often stacking for high resolution); transcription and georeferencing; and also potentially 3D imaging, microscopy, CT scanning and molecular sampling and sequencing, depending upon individual institutional capabilities. Information on the facilities available at each institution is provided on the SYNTHESYS website and is regularly updated. Each institution has appointed a VA Coordinator who will serve as the point of contact for all VA requests during each of the calls and will liaise with both applicants and collections managers in their own institutions. Proposers are expected to work collaboratively with the digitising institution(s) to ensure their proposal meets the criteria. The list of VA Participating Institutions and Coordinator contacts can be found in the Virtual Access section of the SYNTHESYS website.

The 2019 Ideas Call

Because VA is new to the community and to the collections-holding institutions, we decided to invite ideas through an initial Ideas Call, that opened in early October 2019 and closed in late November 2019. This 'soft' call did not offer funding, but provided the opportunity to open discussion between researchers and institutions to shape proposals, preparatory to the first full SYNTHESYS+ VA call in early 2020. We also used this call to test the application process. Applications to the formal call in February 2020 will not be dependent on having participated in the earlier ideas call. Proposals received were examined by the Access Stream Coordinator (S. Knapp), along with other SYNTHESYS+ VA and Management Team members, in order to provide the initial feedback in this Report. We did not formally review or prioritise proposals submitted to the Ideas Call, and proposers did not receive individual feedback, but are expected to use the feedback provided in this report.

Summary of proposals received

Twenty-six complete proposals were received, and more ideas were discussed with VA coordinators but not submitted. Of the submissions to the ideas call:

- Number of proposers varied from one (in 7 proposals) to 15.

- Number of digitising institutions varied from one (in 16 proposals) to 13 (mostly in the range of 2-6 institutions).

- Five proposals submitted indicative cost information, while the remainder did not.

- Proposals were primarily taxonomically focussed, often with a secondary geographic focus and in some instances with a focus on a particular collector.

- One proposal focussed entirely on Paleontological material, and four on both fossil and extant material. The remainder focussed fully on Life Science collections, including a spread across plant (algae, bryophytes, vascular plants) and animal (insects, other invertebrates, vertebrates) groups.

- Proposals covered a wide range of collection and preservation types, including herbarium sheets, envelopes, microscope slides, pinned material, spirit, analog images, and dry material including whole specimens, skeletons and skins.

- Workflows were primarily focused on 2D imaging (often stacking for high resolution), transcription and georeferencing, but also include 3D imaging, microscopy, CT scanning and molecular sampling.

- Five submissions left the field about data licencing blank. Of the 21 that specified potential licences, only two listed exclusively CC-BY-NC, with the majority listing CC0 and CC-BY.

General guidance on VA call proposals

VA is a new approach to collections access, for which proposals need to manage the complexities of working across multiple partners, with varying workflows (and costs), to deliver real impact for the European research community and address Societal Challenges.

Proposals to the ideas call were generally of high quality and potential, reflecting the fact that all of the 19 institutions offering Virtual Access hold collections of global importance.

Following our review of these proposals, the key points that we want to emphasise for proposals to the up-coming Virtual Access Calls are set out below.

Preparation and working with partners

1.1 - It is vital to **contact the VA coordinator(s)** of any institution you may wish to work with as soon as possible, and there will be a deadline for this in Call 1, after which digitising institutions may not be able to support your proposal. Contact with the digitising institutions should be through the coordinator, who will need to promptly engage other relevant staff in their institution (e.g. collections managers, curators) to provide relevant collections information.

1.2 - **Proposal documentation requires the direct input of both the proposer(s) and the digitising institution(s)** (via the VA coordinators). We will make it as clear as possible in the process of Call submissions who is responsible for completing each section (for example digitising institutions should state the timeline for data release); but overall the proposal requires a joint effort and agreement among all parties, and you should allow time for these communications. **Everyone named as a proposer or digitising institution must have been consulted** - please do not include anybody without checking this with them.

1.3 - Proposers should **check for related collections beyond the most obvious one or two** institutions. Even if your proposal picks a small number of collections, it is useful to show why you chose these and what consideration was given to others. For the call, we expect that all VA coordinators will be able to see all proposals, so that they are able to flag additional relevant collections in their institutions.

1.4 - **Digitising institutions are not obliged to support all proposals.** VA coordinators who receive multiple suggested proposals should consider institutional capacity, and which proposals are most likely to be prioritised. Feasibility is also key - digitising institutions should only support proposals where they believe the workflow can be supported successfully (allowing for some being more innovative than others).

1.5 - It is **fine to include proposers/researchers from outside the EU**, and this may strengthen the case for overall impact on societal challenges - **however all proposals must have clear benefit for the European research community**.

Making the best case for your proposal

2.1 - While there is not a hard and fast rule against proposing digitisation of your own institution's collections, or against proposing digitisation of a collection from a single institution, the **most persuasive proposals as a minimum have support from researchers across multiple locations**, and **ideally also include digitisation at more than one collection.** This avoids any perception of benefits being limited to individual researchers or research groups, in line with the overall aspirations of VA to respond to strong demand by current, emerging, and new user communities.

2.2 - **There is a possible exception to this around digitisation of genuinely unique collections**, for instance those that have unique geographic and taxonomic coverage AND offer insights of wider relevance. In such an instance, support from a wide base of researchers remains important, and it is also helpful to set out why collections are unique and what has been done to investigate any related collections.

2.3 - It is also possible that you may choose to limit your proposal to a few digitising institutions e.g. where workflows involve substantial innovation and cost - **coverage should be balanced with feasibility.** If you see your proposal as a first step that may then expand to further collections if successful, please make this clear in your case.

2.4 - While proposals may have many partners, they should have a **single focus area.** If your proposal is linked to others, please state this in the text.

2.5 - It is **vital that proposals are research-enabling**. Workflow innovation is important, and can help make your case for prioritisation, but cannot be the sole or main benefit of a proposal.

2.6 - That said, VA is not a research fund, and **research outputs cannot be funded through these proposals**. Statements about research belong in the impact fields of the submission, not the proposed workflows.

2.7 - When setting out your proposals, it is important to **provide (brief) evidence to support your statements** - for example if you make reference to national or international networks, provide specific details (or ideally engage these in active support of your proposal as co-proposers). Similarly when suggesting research and/or benefits, while these may be forward-looking and not yet under way it is important to **be as specific as possible**. If there is an existing research record in the area covered by your proposal, you may wish to cite these in the impact sections of the submission and say how you expect research to develop.

2.1 - Alongside this, it is useful to state your case as clearly as possible, avoiding or explaining technical terms (such as scientific names of organisms: while all botanists might know what Solanaceae are, some zoologists might not; conversely Heteroptera might be easily understandable to an entomologist, but not to a bryologist). While the prioritisation panel will be expert in our field, their backgrounds and nationalities will vary, and it is **good practice to use language that can be widely and easily understood**.

Data

3.1 - **Both the digitisation AND the data release are to be done by the digitising institution(s)**, not by the proposer(s). This is to ensure efficient and consistent workflows and data. Transnational Access (TA) can include digitisation by individual visitors for their own research: if this is of interest to you, then subject to host approval and support it may be that your proposal is better suited to TA.

3.2 - **Open means open, and should be as fast as possible** - this is a core requirement of Virtual Access and key to the benefits and impact case for proposals. Ideally, data release should take place as projects proceed, or at the latest immediately after completion - reasons for longer timescales should be clearly stated. It is unlikely that projects will be prioritised if data release is subject to legal or other compliance restrictions (e.g. human remains), or is dependent on work outside the scope of the funded project (e.g. would not be released until substantial further work is undertaken).

3.3 - For the ideas call, we offered a choice of licencing types to assess community views on this. The vast majority of ideas call submissions selected non-restrictive licences (CC-0, CC-BY or equivalent), sometimes also selecting Non-Commercial as an option. Only two selected exclusively Non-Commercial licensing, and only one mentioned Share-Alike. Non-commercial and even Share-Alike licences are restrictive in practice (more so than many people realise - for example they limit certain educational uses), so **only non-restrictive licensing will be permitted under the Virtual Access Calls going forward**. This will be clear in the proposal form and guidance.

3.4 - **Data release should be on established institutional or global platforms with longevity, e.g. GBIF, Zenodo, GenBank**. Data release cannot be via citizen science platforms (although they could be used to serve data for uses subsequent to VA projects).

Costs

4.1 - It is clear from this Ideas Call that costing proposals for VA is challenging for all of the institutions involved. In addition, this is a high workload and many proposals were either not ready to undertake detailed costing for the ideas call or did not feel they had sufficient resources to work on costings at this stage. **Further guidance on costs will be provided to VA coordinators and to proposers** (see below).

4.2 - Costs will be allocated to institutions on a demand basis in line with the prioritised VA proposals, regardless of what was in the original institutional indicative bids in the SYNTHESYS+grant proposal (Smith et al. 2019).

4.3 - In very broad terms, allowable costs are those categories that institutions included in their original indicative VA bids, and which go towards creating Virtual Access to data (including images etc). We **expect that the majority of costs will be for staff resources, including digitisers, and for consumables such as barcodes**. **No capital costs** e.g. for lasting and substantial equipment are allowable. No research costs are allowable. Costs for preparation e.g. curator time, and for data release e.g. limited amounts of development, may be allowable depending on their place in prioritised proposals. Major database development is unlikely to be allowable - **data release should make use of existing databases and websites** (e.g. Collections management systems, aggregators and institutional sites).

Feedback on the process

Overall, Ideas Call submissions were of high quality and addressed or began to address the relevant criteria. Submissions were diverse, and some raised concerns about feasibility; proposers will need to be aware that **not all digitisation-on-demand through this programme will be feasible at all participating institutions**. Early and on-going communication with VA coordinators will be critical to the development of successful proposals. An indicative flow for proposals is outlined in Fig. 1; this is an outline schema and not all communications and activities are necessarily depicted.

Figure 1. doi

Indicative proposal flow for SYNTHESYS+ Virtual Access proposals. Blue boxes indicated proposer(s) actions/responsibilities; violet boxes VA coordinator actions/responsibilities; green boxes indicate where communication and discussion between proposers and VA coordinators will be paramount; white (no fill) boxes indicate post-proposal actions. Double-ended arrows indicate where back-and-forth will occur between proposer(s) and VA coordinator(s).

Word counts or **word limits** will be included in the submission process, to provide guidance on the appropriate level of detail for text fields and avoid a high degree of variety between very short and very long submissions.

It is clear that **costs** are very challenging (see also comments on this above). For the Call, we will ensure that:

- The template allows both for each digitising institution to complete separate cost details, and for a total of costs for the project to be provided.
- Guidance is provided on what cost categories can and can't usually be included, as well as on how to get further information e.g. from the institutions offering Virtual Access.
- Cost templates allow for costs at various stages of potential workflows and/or can be tailored by institutions (current templates only cover imaging)

We encourage VA coordinators / digitising institutions to **consider calculating and agreeing costs for various stages and workflows prior to the Call** if possible, particularly where there are existing workflows, so that these can be provided and used consistently across any proposals received.

Digitising institutions will need to consider the balance that best suits them between hiring new resources and using existing resources to work on funded proposals - **if hiring is necessary, timing for this should be built into proposal timing**.

The **process has been quite time consuming for Proposers and VA coordinators** in many instances. To some degree this is unavoidable with a new process, however The NHM Management Office will work with VA coordinators to minimise this where possible. In particular, we will try to ensure that Coordinators do not need to provide support with completing the submission process (e.g. what goes in what sections or how to manage online submission), but can focus on input to the relevant fields for their institution.

Funding program

H2020-INFRAIA-2018-2020 (Integrating and opening research infrastructures of European interest) Topic: INFRAIA-01-2018-2019

Grant title

Synthesis of Systematic Resources (SYNTHESYS PLUS)

Hosting institution

Natural History Museum, London

Ethics and security

No ethics and security issues to declare

Author contributions

Helen Hardy and Sandra Knapp wrote the initial draft report and are joint first authors; all other authors contributed to the text of the final version (names listed in alphabetical order).

Conflicts of interest

All authors declare no conflicts of interest.

References

- Addink W, Koureas D, Rubio A (2019) DiSSCo as a New Regional Model for Scientific Collections in Europe. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37502
- Lannom L, Koureas D, Hardisty A (2019) FAIR Data and Services in Biodiversity Science and Geoscience. Data Intelligence 122-130. https://doi.org/10.1162/dint_a_00034
- Smith V, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston E, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e46404