SYNTHESYS+ VA Call 2 Review

Review Board: Arturo H. Ariño, Joseph T. Miller, Pamela S. Soltis

Preface:
The Review Board congratulates the SYNTHESYS+ VA Team on developing and implementing a program that has catalyzed the digitization process across institutions and across national borders. The results of the two rounds of VA Calls have generated substantial resources for the research community while developing and providing technical expertise and interinstitutional collaboration. This Review focuses on the questions raised in our Charge.

Review Board to comment on these items:

1. Is the process as it worked in Call 1 fair and unbiased?*
2. Does the process we followed fulfil the brief in the SYNTHESYS+ grant agreement?*
3. **Have the data been published in a FAIR and open manner?**

Some data have been published, but even after two years since the completion of the projects, much of the VA1 data is not published, or only published at home institutions. In no case is the complete set of collections, from multiple institutions, available in one place. We should distinguish two separate facets to this issue: (1) whether data have been published, and (2) whether published data are FAIR and open: while portals to access the metadata may have been created, data may not have been linked yet at all, so they technically would not be published and therefore not FAIR (FAIR requires publication as a prerequisite), or if available, they might be behind a gate with limited provision for reuse (e.g. requiring manual searches or lacking API), so they might be "published" but not fully FAIR.

In the cases reviewed, some data from the majority of projects (⅘) are published by at least some partners in the project, but their FAIRness is variable. Three are mostly FAIR inasmuch as at least one FAIR-compliant gateway (GBIF) is surrogated. One case used its own portal (NHM) with API available. We found no published data about one project.

Various delays – such as COVID, co-development of ELVIS during the digitization period for VA1 – undoubtedly contributed to some of the delays in making data from VA1 FAIR and open. However, even the data that have been made 'public' are not necessarily Findable. Specimens shared via an institutional website, in and amongst all other records, are not at all Findable with other digitized data for the same SYNTHESYS+ VA project, nor are the data Reusable as a data set (or even the basis of a data set) because it would be onerous (perhaps impossible) to pull the data from multiple portals together by a potential user. It is stated that much of the VA2 material will be public in August or September 2023. The Kranz VA2 material in Lumous Finland is published to GBIF and has 15 citations. Failure to make data FAIR is a lost opportunity for VA-funded digitization to be further reused.

In several cases, we found that VA2-digitized data may have been published in a FAIR repository, but lack of adequate metadata linking these published data to the VA2 effort makes discovery (as a VA "collection") nearly impossible and prevents monitoring its impact. The case affects all projects and 47% of all records. Therefore, only 26% of the digitized effort could at most be considered, at the time of writing, FAIR data (but more may become FAIR in the near future).

On a project-by-project and partner-by-partner basis, data availability and FAIRness become

very variable (Table 1). See individual projects for additional details at the end of this report.

Table 1. Summary status of projects as of Aug. 28th, 2023.

Project	Data portals	Data digitized	Data published	Published data open/FAIR
1-Wheat	https://data.nhm.ac.uk/dataset/wheat-through-the-ages/resource/04844d1b-9dcc-4a2a-b6a9-cb930e30e6e2 (July 23rd) for NHM records. HUJI, RBGK not available.	Yes (136%, 83%-175%). Holdings underestimated (NHM).	Partially (64%). 12+K records by NHM.	Some, through NHM data portal. API available.
2-BIT	https://digital.csic.es/handle/10261/307317 (May 18th) for MNCN records. Also through GeoCASE and GBIF.	Yes (463%, 17% - 717%). Holdings underestimated (NHM).	Partially (8%). 1+K records by MNCN and NHMW	Yes, through GeoCASE and GBIF. API available.
7-Krantz	Uses GBIF	Yes (94%), most partners 100%	Partially (80%). Some expected October 2023+.	Some, though GBIF (API available) but most lack metadata.
8-Xenopus	Not found	Yes (90%)	Not found	Not found
10-Cyrtandra	Uses JACQ's and GBIF	Yes (88%)	Partially (55%)	Some, through GBIF (API available) but most lack metadata.

4. Is there any evidence that the community is using these data to advance science, even given the early stage of the release of these data?

There is very little evidence that the community is using these data at this point, which is perhaps not surprising. 50% of the digitized data have not yet been published. Of those that are, most of the data digitized through these projects have just recently been released through or aggregated to data portals (in most cases merely a few weeks or months before this report), or are not yet released. Multi-year lags are known to exist between digitization and publication (Gaijy et al., 2013), and it can be expected that similar lags exist between data availability and discovery for purpose by third parties, which add to the typical research publication lag. Assessment of community intake should likely be done with at least one year delay after data publication–earlier uptake might more likely be achieved by the group involved in the digitization or prompting it. In addition, one should distinguish the use of *the digitized data* from the use of the data *after being digitized*. In some projects, historical records dating back several decades have been digitized, but such data could have been used in earlier research, e.g. in the original papers associated with the expedition whence they came. In our assessment, we focus only on digitized data after becoming publicly available in a FAIR repository, and therefore the time span (and opportunity for uptake by the community) is much more limited.

5. Did we have adequate reach to requesters, if not how could this be improved?

It appears that the solicitation was widely viewed. The receipt of 32 VA requests is substantial and indicates that the advertisement reach was sufficient. This is further supported in that all 20 participating SYNTHESYS+ VA partners were involved in the calls. Requesters also took advantage of the ability to request digitization in more than one collection.

6. Does the balance of requests show evidence of reaching new communities of users?

The types of collections and digitization projects were impressively broad, including botany, entomology, paleontological, anthropological, and aerial image collections. However, it was somewhat difficult for the Review Board to evaluate whether or not these were new user communities without knowing the extent of digitization in these communities prior to these requests. The Review Board members were surprised to see DNA sequencing expenses included, but sequencing was noted as part of digitization as defined by the VA call. Development of the cost and quality control structure for inclusion of sequencing must have brought its own challenges, but this extension of digitization to the community of users interested in sequence data appears new to VA2. Was this work successful? No mention of submission of DNA sequence data to repositories was noted in the Summary.

7. Was the balance of requesters wide enough geographically? Suggestions for ways in which could be improved/expanded for the programme going forward as part of DiSSCo?

Thirty-two requests were received, with data requested from all 20 participating institutions and from 12 countries (Figure 1). There was a reasonable linear bias (neither too logarithmic, indicative of resource hoarding, or uniform, where countries with numerous institutions could be negatively biased). GBR (53%) and DEU (47%) were in the consortia for half the projects through one institution or other (GBR: 3 institutions, DEU: 5 institutions). BEL was represented by 3 institutions in one-third of the projects. FIN and SWE were in the least projects (9%) and had the lowest project/institution ratio (3 projects/institution) while AUT had the highest ratio (12 projects/institution).

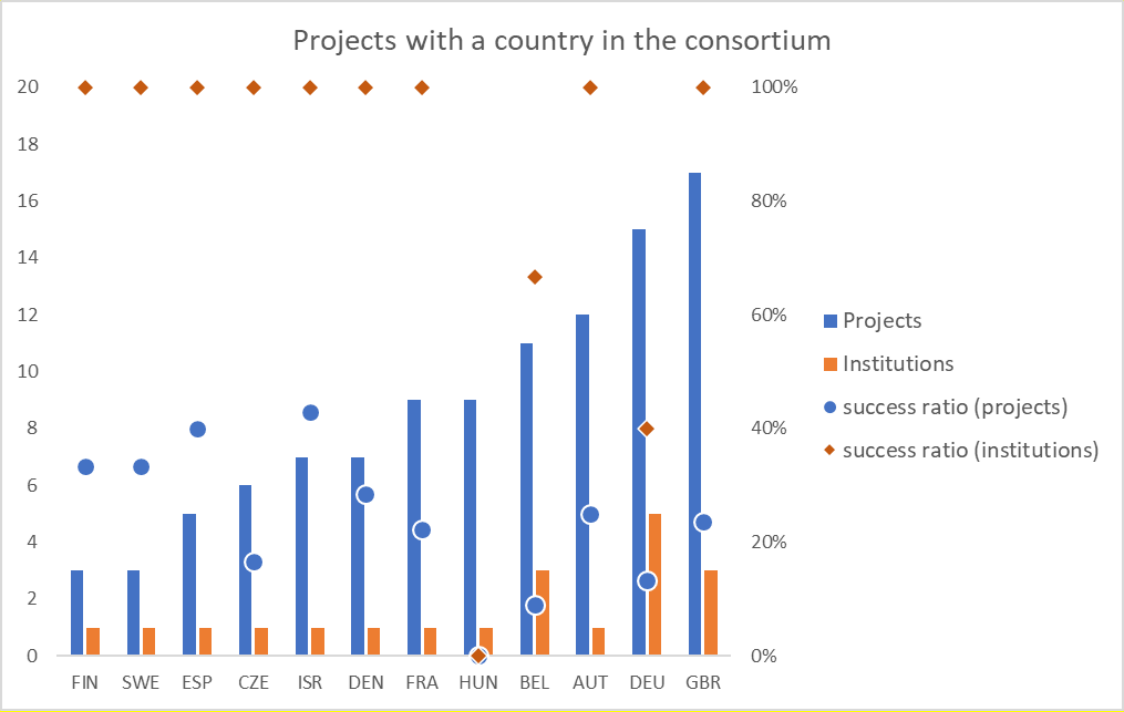

Figure 1: Distribution of requested projects and institutions by country and relative success ratio.

8. Any comments on the VA Coordinator role? Single vs. shared role, scope of work?*

9. Any other feedback you might want to give us to help improve VA going into the future as the SYNTHESYS effort ends and the DiSSCo consortia take responsibility for this activity. In particular any suggestions for a balance/preference between national and European level VA activities would be extremely helpful.

A great way to showcase the value of VA projects is to be able to access all specimen information from a particular project at a single place, for example, having the 38,000 *Dianthus* digitized specimens from the seven collections on a single data set in a portal. Apart from the SYNTHESYS+ project, GBIF has developed hosted portals that can provide this along with citation and usage tracking. This is a type of a funders view of data and would be valuable to show the EC the added value of specific digitization projects it funded. A VA-hosted portal could be considered now as data are being tagged and published to GBIF.

Unfortunately, this is generally not yet happening in VA2 (Table 2). Some projects actually seemed to have their data published as FAIR, and some had metadata published, enabling linking the data to the digitizing effort. But in many cases, even though digitized specimens were reported as uploaded to a FAIR repository, lack of reference to the digitizing project prevented them from being linked to this digitized “collection”, i.e. information about the role of VA was apparently lost (or could not be located through searches or calls).

Table 2. Summary achievement of goals as of Aug. 28th, 2023

	01-Wheat			02-BIT			07-Krantz			08-Xenopus			10-Cyrtandra		
Institution	Dig.	Pub.	FAIR	Dig.	Pub.	FAIR	Dig.	Pub.	FAIR	Dig.	Pub.	FAIR	Dig.	Pub.	FAIR
HUJI	83%	0%	0%	17%	0%	0%	90%	0%	0%						
LUOMUS							100%	100%	0%						
MfN				100%	0%	0%	126%	0%	0%						
MNCN				527%	100%	100%	100%	0%	0%						
MNHN				52%	100%	0%	100%	91%	0%						
NHM	175%	90%	100%	717%	0%	0%	100%	100%	100%						
NHMW				75%	100%	0%	110%	100%	0%				101%	100%	0%
NMP							0%		0%						
NRM							67%	0%	0%						
RBGE													103%	100%	0%
RBGK	93%	2%	0%										78%	15%	0%
RBINS										83%	0%	0%			
RMCA										100%	0%	0%			
SMNS							43%	0%	0%						
UCPH													100%	0%	0%
Total	136%	64%	99%	463%	8%	58%	94%	80%	14%	90%	0%	0%	88%	55%	0%

**FAIR prerequisite is publication so this fraction is over published data.*

It should be noted that of 26 partner-project tasks, 10 expected to publish no sooner than August 2023 (the date of this report), and a further 10 had set July as publication date. Even slight delays may therefore account for well one-third of unpublished data. We believe that the fraction of published data, and hopefully of FAIR data, will grow significantly in the coming weeks or months.

Although there was concern about a possible risk of running VA and TA concurrently such that the VA Call would be missed by research communities who would have otherwise benefited from it, this seems not to have been an impediment. The communication strategy that included broad publicity about the Calls and coordination of internal stakeholders through the VA Coordinators from each SYNTHESYS+ institution seems to have been sufficient, with advertisements by email, social media, and project websites. The fact that 32 proposals were received argues that potential users were aware of the VA Call.

With a goal of eventually digitizing all collections, DiSSCo will want to balance digitization at national vs. European scales. Prioritizing specimens requested for specific data use, as with the SYNTHESYS+ VA program will likely lead to more rapid use by the research community; thus, a program that extends across institutions and countries will yield greater impact in less time. However, although balancing the need for access to large amounts of digitized data for research with the potential benefits of digitizing in ALL collections (such as exposing new data, benefits of local resources and information on rare species, enhanced training, etc.) is a challenge, DiSSCo will need to consider the digitization landscape holistically and design calls and make awards based on a diversity of criteria.

Review of Funded Proposals:
#1- Wheat Through the Ages

All data reported as digitized. NHM data are searchable through NHM general data portal since

July 23rd. The full dataset can be obtained by filtering through the project's name as a CSV file, and a Darwin Core-compliant version can be generated. Individual records can be served via API. Therefore, data can be considered reasonably FAIR even though not integrated with other aggregators. Data from HUJI and RBGK not found (RBGK expected next month).

#2 Bryozoa Identification Tool (BIT) For Quaternary and Recent Mediterranean and North Atlantic Bryozoans

The digitization target has been met, with data published May 18th for MNCN images available; index available and actionable. In addition to its own portal serving data and images, data have been uploaded and integrated to a FAIR-compliant infrastructure (GBIF). Data from NHM might be published specimen-by-specimen but metadata about the project cannot be found for verification. HUJI publication date not set.

#7 Harmonizing verbatim names in digitized collections – the Krantz material as a model

This is the project having the most partners (10). Most digitization has been completed, but data from one-half of the partners are not yet available. Although published data from the other half seem to account for 80% of the expected digitization volume and metadata for some collections exist, the actual data are not readily findable for want of reference to the project, giving it a low fraction of FAIR-identified data.

#8 Monitoring Climate Change, Environmental Pressure on biodiversity and Invasiveness using Xenopus as a model system

The digitization target has nearly been met, but no data are yet available.

#10 Accelerating taxonomic progress on the large rainforest genus Cyrtandra

Digitization is nearly complete (about 500 short of target); full data set from RBGE and NHMW available, not available from RBGK, but with target of August.

**These were covered in the 1st review, and there has been no significant change to the process followed nor to the VA Coordinator role.*