Research Ideas and Outcomes 5: e46404
doi: 10.3897/rio.5.e46404

Grant Proposal

SYNTHESYS+ Abridged Grant Proposal

Vincent Stuart Smith[‡], Kristina Gorman[‡], Wouter Addink[§,|,¶], Christos Arvanitidis[#], Ana Casino[¤], Katherine Dixey[‡], Gabriele Dröge[«], Quentin Groom[»], Elspeth Margaret Haston[^], Donald Hobern[˅,¦], Sandra Knapp[‡], Dimitrios Koureas[§,|], Laurence Livermore[‡], Ole Seberg[ʔ]

‡ The Natural History Museum, London, United Kingdom
§ Naturalis Biodiversity Center, Leiden, Netherlands
| Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
¶ Species 2000 Secretariat, Leiden, Netherlands
# Hellenic Centre for Marine Research, Heraklion, Greece
¤ Consortium of European Taxonomic Facilities, Brussels, Belgium
« Freie Universität Berlin, Botanic Garden and Botanical Museum, Berlin, Germany
» Meise Botanic Garden, Meise, Belgium
^ Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom
˅ Global Biodiversity Information Facility, Copenhagen, Denmark
¦ International Barcode of Life, Canberra, Australia
ʔ Natural History Museum of Denmark, Copenhagen, Denmark

Corresponding author: Vincent Stuart Smith (vince@vsmith.info)

Reviewable v1

Received: 09 Sep 2019 | Published: 09 Sep 2019

Citation: Smith VS, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston EM, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5: e46404. https://doi.org/10.3897/rio.5.e46404

Abstract

European natural history collections are a critical infrastructure for meeting the most important challenge humans face over the next 30 years – creating a sustainable future for ourselves and the natural systems on which we depend – and for answering fundamental scientific questions about ecological, evolutionary, and geological processes. Since 2004 SYNTHESYS has been an essential instrument supporting this community, underpinning new ways to access and exploit collections, harmonising policy and providing significant new insights for thousands of researchers, while fostering the development of new approaches to face urgent societal challenges. SYNTHESYS+ is a fourth iteration of this programme, and represents a step change in the evolution of this community. For the first time SYNTHESYS+ brings together the European branches of the global natural science organisations (GBIF https://www.gbif.org/, TDWG https://www.tdwg.org/, GGBN http://www.ggbn.org/ggbn_portal/ and CETAF https://cetaf.org/) with an unprecedented number

of collections, to integrate, innovate and internationalise our efforts within the global scientific collections community. Major new developments addressed by SYNTHESYS+ include the delivery of a new virtual access programme, providing digitisation on demand services to a significantly expanded user community; the construction of a European Loans and Visits System (ELViS) providing, for the first time, a unified gateway to accessing digital, physical and molecular collections; and a new data processing platform (the Specimen Data Refinery), applying cutting edge artificial intelligence to dramatically speed up the digital mobilisation of natural history collections. The activities of SYNTHESYS+ form a critical dependency for DiSSCo - the Distributed System of Scientific Collections (https://dissco.eu/), which is the European Research Infrastructure for natural science collections, under the ESFRI umbrella. DiSSCo will undertake the maintenance and sustainability of SYNTHESYS+ products at the end of the programme.

Keywords

natural history collections, research infrastructure, global natural science, digitisation, collaboration, data standards, DiSSCo

Context

The world's natural science collections are an extraordinary data vault – a vast mine of unseen and under-exploited information on the natural world, of critical relevance to addressing many of the big challenges facing science and society. For the past 15 years a consortium of natural science museums, botanic gardens, universities and businesses, led by the Natural History Museum, London, have coordinated a major project called SYNTHESYS (Synthesis of Systematic Resources) to improve and expand access to natural science collections across Europe. SYNTHESYS, alongside related communities such as CETAF, TDWG and GBIF has been central to cross European efforts to structure activities, building a network of researchers and projects to unlock data associated with the hundreds of millions of specimens in European collections. Supporting more than 4,000 research projects, generating 4,500 publications, and aligning the policies, processes and ambitions of the European natural sciences collections community, SYNTHESYS has to date, had a transformative impact not only on the work of these institutions but also the wider research community, impacting topics as diverse as climate change and conservation to human health and agriculture.

In February 2019 this group launched SYNTHESYS+ (submitted as SYNTHESYS PLUS), the most ambitious iteration of the project so far. SYNTHESYS+ was a response to Call H2020-INFRAIA-2018-1 from the Research and Innovation Directorate of the European Commission of the European Commission and scored 14.5 of a possible 15 points by the European commission's external review panel. With an increasing focus on digital collections, SYNTHESYS+ is pioneering the first pan-European programme of digital access to specimens, allowing researchers to request digitisation and the development of

the mass digitisation workflows necessary to support this effort. SYNTHESYS+ is also underpinning developments central to DiSSCo - the Distributed System of Scientific Collections, which is a successor programme involving more than 115 natural Science collections across Europe.

This paper is an abridged version of the original proposal. It contains the overarching scientific case for SYNTHESYS+, alongside a description of our major activities. Differences between this paper and the full “Description of Work” include redactions of financial and personal information alongside our risk analysis; inclusion of additional citations that could not be included in the original proposal due to space limitations; minor edits to improve readability; and the inclusion of higher resolution versions of the figures. The abridged proposal is published here to frame the publication of future outputs from SYNTHESYS+, and to serve as a potential model for other global efforts to help coordinate and strengthen related activities. Across the globe, a series of major programmes are underway to unlocking the estimated 1.5 Billion specimens contained in natural science collections across the world. We hope SYNTHESYS+ will inspire the next generation of cross-continental activities, strengthening coordination as we work toward a common global infrastructure to support access and research on the natural science collections of the world.

List of beneficiaries

1. **Natural History Museum (NHM)**†, established in CROMWELL ROAD, LONDON SW7 5BD, United Kingdom
2. **Naturhistorisches Museum Wien (NHMW)**, established in BURGRING 7, WIEN 1010, Austria
3. **Institut royal des Sciences naturelles de Belgique (RBINS)**, established in RUE VAUTIER 29, BRUXELLES 1000, Belgium
4. **Musée royal de l'Afrique centrale (RMCA)**, established in LEUVENSESTEENWEG 13, TERVUREN 3080, Belgium
5. **Agentschap Plantentuin Meise (BGM)**, established in NIEUWELAAN 38, MEISE 1860, Belgium
6. **Consortium of European Taxonomic Facilities (CETAF)**, established in RUE VAUTIER 29, BRUXELLES 1000, Belgium
7. **Národní muzeum-National Museum NM (NMP)**, established in VACLAVSKE NAM 68, PRAHA 11579, Czech Republic
8. **Freie Universität Berlin (BGBM)**, established in KAISERSWERTHER STRASSE 16-18, BERLIN 14195, Germany
9. **Museum für Naturkunde - Leibniz-Institut für Evolutions- und Biodiversitätsforschung an der Humboldt-Universität zu Berlin (MFN)**, established in INVALIDENSTRASSE 43, BERLIN 10115, Germany
10. **Senckenberg Gesellschaft für Naturforschung (SGN)**, established in SENCKENBERGANLAGE 25, FRANKFURT 60325, Germany
11. **Staatliches Museum für Naturkunde Stuttgart (SMNS)**, established in ROSENSTEIN 1, STUTTGART 70191, Germany

12. **Zoologisches Forschungsmuseum Alexander Koenig (ZFMK)**, established in ADENAUERALLEE 160, BONN 53113, Germany
13. **Københavns Universitet (UCPH)**, established in NORREGADE 10, KOBENHAVN 1165, Denmark
14. **Global Biodiversity Information Facility (GBIF)**, established in Universitetsparken 15, Copenhagen 2100, Denmark
15. **Agencia Estatal Consejo Superior de Investigaciones Científicas (CSIC)**, established in CALLE SERRANO 117, MADRID 28006, Spain
16. **Helsingin yliopisto (LUOMUS)**, established in FABIANINKATU 33, HELSINGIN YLIOPISTO 00014, Finland
17. **Muséum national d'Histoire naturelle (MNHN)**, established in RUE CUVIER 57, PARIS 75005, France
18. **A.2.I.A. ANALYSE D IMAGE ET INTELLIGENCE ARTIFICIELLE - ARTIFICIAL INTELLIGENCE AND IMAGE ANALYSIS SA (A2iA)**[*], established in RUE DE LA BIENFAISANCE 37-39, PARIS 75008, France
19. **Hellenic Centre for Marine Research (HCMR)**, established in LEOFOROS ATHENS SOUNIO 46 7KM, ATTIKIA ANAVISSOS 19013, Greece
20. **ΕΘΝΙΚΟ ΔΙΚΤΥΟ ΕΡΕΥΝΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΑΕ / Ethniko Diktyo Erevnas Technologias A.E. (GRNET)**, established in LEOFOROS KIFISIAS 7, ATHINA 11523, Greece
21. **Magyar Természettudományi Múzeum (HNHM)**, established in BAROSS UTCA 13, BUDAPEST 1088, Hungary
22. **The Hebrew University of Jerusalem - האוניברסיטה העברית בירושלים (HUJI)**, established in GIVAT RAM CAMPUS, JERUSALEM 91904, Israel
23. **Stichting Naturalis Biodiversity Center (Naturalis)**, established in DARWINWEG 2, LEIDEN 2333CR, Netherlands
24. **Picturae BV (Picturae)**, established in DE DROOGMAKERIJ 12, HEILOO 1851 LX, Netherlands
25. **Stichting International Working Group on Taxonomic Databases (TDWG Europe)**, established in ZANDHEUVEL 52B, OOSTERHOUT 4901 HW, Netherlands
26. **Naturhistoriska riksmuseet(NRM)**, established in Frescativägen 40, STOCKHOLM SE 114 18, Sweden
27. **Göteborgs universitet (UGOT-GGBC)**, established in VASAPARKEN, GOETEBORG 405 30, Sweden
28. **Royal Botanic Garden Edinburgh (RBGE)**, established in INVERLEITH ROW 20A, EDINBURGH EH3 5LR, United Kingdom
29. **Royal Botanic Gardens, Kew (RBGK)**, established in ROYAL BOTANIC GARDENS KEW, RICHMOND TW93AB, United Kingdom
30. **DIGIRATI LIMITED (Digirati)**, established in THE HUB, 70 PACIFIC QUAY, GLASGOW G51 1EA, United Kingdom
31. **The University of Manchester (UNIMAN)**, established in OXFORD ROAD, MANCHESTER M13 9PL, United Kingdom
32. **Smithsonian Institution (SmithsonianGGBN)**[‡], established in JEFFERSON DRIVE 1000, WASHINGTON DC 20560, United States, as ‘beneficiary not receiving EU funding’

† The project coordinator, *this project partner formally withdrew from the project in February 2019, ‡ beneficiary not receiving EU funding. Data also available in Table 3.

Table 1.

New user communities

Environment	Agriculture	Health	Border control	Biobanking
• Urban planning • Environmental • impact assessment • Deep-sea mining • Conservation planning & monitoring	• Species identification • Future domestication • Land use change • Industrial (insect) farming • Forestry	• Pathogen identification • Medicine and food supplement verification • Pharmaceutical industry	• Invasive species and pests • CITES protected species enforcement • Countering illegal wildlife trade identification	Preserve geneticmaterial (tissues & seeds) for: • Research • Government • Industry (medicine, biotech. & agriculture)

Table 2.

SYNTHESYS+ at-a-glance summary of work packages within their respective streams.

Networking Activities Stream (1.3.1)	Joint Research Activities Stream (1.3.2)	Access Stream (1.3.3)
NA1: Management Overseeing the coordination (including Stream coordination) and management of SYNTHESYS+ **NA2: Harmonisation of policies, best practices, training & support** Supporting alignment of best practice in policy and delivery of best practice in training. **NA3: Collections in the age of genomics - standards & processes** Development, implementation and dissemination of standardised best practices tosupport molecular sequencing and biobanking activities. **NA4: Digital Standards & Processes** Increasing standards compliance across the community to improve interoperabilitysupporting digital access to collections facilitating Digitisation on Demand (DoD). **NA5: Internationalisation and engagement of new user communities in EC priority areas** Integration of major international stakeholders& international data gathering to develop theglobal collections research agenda.	**JRA1: Optimisation of Access** Developing an integrated system (ELViS) to support loans, visits, physical / digital access requests & track outputs. **JRA2: Collections on Demand** Prioritising VA requests and developing the technologies to fill gaps in our institutional capacity to undertake DoD and sequencing on demand. **JRA3: Specimen Data Refinery** Developing the community platform and processes to extract, enhance andannotate data from digital images and records, using artificial intelligence(machine learning & computer vision) and human-in-the loop approaches.	**TA1: Physical Transnational Access** Expanding physical Access to an enhanced network of institutions to reach new user communities, including targeting priority user groups and addressing societal challenges. **VA1: Virtual Access** Enabling users to request access to digital surrogatesacross multiple collections through DoD workflows, with open data deposited and freely accessible in internationally recognised repositories operating FAIR data standards.Includes: • DoD support • Open data via Data Portals • Data registration, tracking & citation

Table 3.

Members of the consortium. †The project coordinator, *this project partner formally withdrew from the project in February 2019, ‡ beneficiary not receiving EU funding

Consortium member	Abbreviation	Address	Country
Natural History Museum†	NHM	CROMWELL ROAD, LONDON SW7 5BD	United Kingdom
Naturhistorisches Museum Wien	NHMW	BURGRING 7, WIEN 1010	Austria
Institut royal des Sciences naturelles de Belgique	RBINS	RUE VAUTIER 29, BRUXELLES 1000	Belgium
Musée royal de l'Afrique centrale	RMCA	LEUVENSESTEENWEG 13, TERVUREN 3080	Belgium
Agentschap Plantentuin Meise	BGM	NIEUWELAAN 38, MEISE 1860	Belgium
Consortium of European Taxonomic Facilities	CETAF	RUE VAUTIER 29, BRUXELLES 1000	Belgium
Národní muzeum-National Museum NM	NMP	VACLAVSKE NAM 68, PRAHA 11579	Czech Republic
Freie Universität Berlin	BGBM	KAISERSWERTHER STRASSE 16-18, BERLIN 14195	Germany
Museum für Naturkunde - Leibniz-Institut für Evolutions- und Biodiversitätsforschung an der Humboldt-Universität zu Berlin	MFN	INVALIDENSTRASSE 43, BERLIN 10115	Germany
Senckenberg Gesellschaft für Naturforschung	SGN	SENCKENBERGANLAGE 25, FRANKFURT 60325	Germany
Staatliches Museum für Naturkunde Stuttgart	SMNS	ROSENSTEIN 1, STUTTGART 70191	Germany
Zoologisches Forschungsmuseum Alexander Koenig	ZFMK	ADENAUERALLEE 160, BONN 53113	Germany
Københavns Universitet	UCPH	NORREGADE 10, KOBENHAVN 1165	Denmark
Global Biodiversity Information Facility	GBIF	Universitetsparken 15, Copenhagen 2100	Denmark
Agencia Estatal Consejo Superior de Investigaciones Científicas	CSIC	CALLE SERRANO 117, MADRID 28006	Spain
Helsingin yliopisto	LUOMUS	FABIANINKATU 33, HELSINGIN YLIOPISTO 00014	Finland
Muséum national d'Histoire naturelle	MNHN	RUE CUVIER 57, PARIS 75005	France

Consortium member	Abbreviation	Address	Country
A.2.I.A. ANALYSE D IMAGE ET INTELLIGENCE ARTIFICIELLE - ARTIFICIAL INTELLIGENCE AND IMAGE ANALYSIS SA*	A2iA	RUE DE LA BIENFAISANCE 37-39, PARIS 75008	France
Hellenic Centre for Marine Research	HCMR	LEOFOROS ATHENS SOUNIO 46 7KM, ATTIKIA ANAVISSOS 19013	Greece
ΕΘΝΙΚΟ ΔΙΚΤΥΟ ΕΡΕΥΝΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΑΕ / Ethniko Diktyo Erevnas Technologias A.E.	GRNET	LEOFOROS KIFISIAS 7, ATHINA 11523	Greece
Magyar Természettudományi Múzeum	HNHM	BAROSS UTCA 13, BUDAPEST 1088	Hungary
The Hebrew University of Jerusalem - האוניברסיטה העברית בירושלים	HUJI	GIVAT RAM CAMPUS, JERUSALEM 91904	Israel
Stichting Naturalis Biodiversity Center	Naturalis	DARWINWEG 2, LEIDEN 2333CR	Netherlands
Picturae BV	Picturae	DE DROOGMAKERIJ 12, HEILOO 1851 LX	Netherlands
Stichting International Working Group on Taxonomic Databases	TDWG Europe	ZANDHEUVEL 52B, OOSTERHOUT 4901 HW	Netherlands
Naturhistoriska riksmuseet	NRM	Frescativägen 40, STOCKHOLM SE 114 18	Sweden
Göteborgs universitet	UGOT-GGBC	VASAPARKEN, GOETEBORG 405 30	Sweden
Royal Botanic Garden Edinburgh	RBGE	INVERLEITH ROW 20A, EDINBURGH EH3 5LR	United Kingdom
Royal Botanic Gardens, Kew	RBGK	ROYAL BOTANIC GARDENS KEW, RICHMOND TW9 3AB	United Kingdom
DIGIRATI LIMITED	Digirati	THE HUB, 70 PACIFIC QUAY, GLASGOW G51 1EA	United Kingdom
The University of Manchester	UNIMAN	OXFORD ROAD, MANCHESTER M13 9PL	United Kingdom
Smithsonian Institution‡	SmithsonianGGBN	JEFFERSON DRIVE 1000, WASHINGTON DC 20560	United States

1. Excellence

1.1 Objectives

Summary

The world's natural history (NH) collections contain at least 2 billion specimens, of which an estimated 55% reside in Europe (Ariño 2010). The European collections are a major research resource representing 80% of the World's bio- and geo-diversity. Collected over 500 years of human exploration and still constantly expanding, these are a vast and underused repository of information about the natural world. They provide a unique data source and research tool to address the most important challenge that humans face over the next 30 years - creating a sustainable future for ourselves and the natural systems on which we depend and for answering fundamental scientific questions about key ecological, evolutionary, and geological processes and how they interact to shape our changing planet (Suarez and Tsutsui 2004).

SYNTHESYS+ aims to unify operations and access for European natural science collections. The longevity and distributed nature of biodiversity collections has inevitably led to divergence in the procedures and standards, even though overall goals of these organisations are well aligned. SYNTHESYS+ is designed to address this problem. The workplan will transform the community into an integrated, data-driven, pan-European research infrastructure (RI), organising the information about the natural environment contained in these collections, and make it universally accessible and useful to researchers tackling scientific and societal challenges and to the wider community.

The overall objectives of SYNTHESYS+ are to:

1. **transform the fragmented access model to a central facility and support new forms of virtual access within an integrated, data-driven, pan-European research infrastructure**;
2. **coordinate formal and professional training activities to enhance data skills for scientists**;
3. **run a robust research programme, enabling scientists to benefit from new digital and genomic RIs to deliver data-driven scientific innovation**;
4. **develop common policies, harmonising digital and molecular processes in alignment to national and international standards**;
5. **work towards the establishment of a shared international vision of bringing together all global natural science collections as an integrated RI.**

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+ acts as a fourth iteration of this programme as it evolves into a sustainable and independent RI through the DiSSCo (Distributed System of Scientific Collections) ESFRI initiative. **As the volume and diversity of information derived from natural science collections exponentially increase, so does the need for infrastructures that provide access to large volumes of linked and precise data from these collections. SYNTHESYS+ brings an unprecedented number of collections institutions together with the European branches of the global natural science organisations to address this challenge.**

Background

Data derived from European natural science collections underpin countless discoveries and innovations, including tens of thousands of scholarly publications and official reports annually (used to support legislative and regulatory processes relating to health, food, security, sustainability and environmental change); inventions and products critical to our bio-economy; databases, maps and descriptions of scientific observations; instructional material for students, as well as educational material for the public.

In the last decades, research practice has changed dramatically. Remote sensing, rapid identification and molecular approaches allow us to efficiently monitor the changing world and to better understand the causes of those changes (Kelling et al. 2009; Shokralla et al. 2012). As the volume and diversity of information derived from natural science collections is exponentially increasing, so is the need for adequate infrastructures that go further than providing simple access to different data classes. A holistic approach is now required (Fig. 1) to effectively underpin the entire research lifecycle and provide open access to mass, linked and precise data (Hardisty et al. 2013).

New technologies are providing opportunities to combine the data held in NH collections with other sources on species, genomes, phenotypes, geography, and the environment in ways that drive novel, integrative research. Prime examples of this are (1) the huge compilation of data on the distribution of living species that is held by the Global Biodiversity Informatics Facility (GBIF), (2) the genetic sequence information that is collated by GenBank and iBOL and (3) the data on morphology held by MorphoBank and TraitBank. NH specimens are the primary link between disparate biological and geological data. **Mobilising these data through digitisation and genomic approaches, coupled with linking these datasets across domains, provides an incredibly powerful research tool for understanding the past, present and future of the natural world.**

At present, however, the **exploitation of these opportunities is severely limited by the low proportion of the collections that are digitally accessible; the lack of a common platform for access to NH specimen information; incomplete links between major data sources about the natural world; and weak informatics tools to facilitate data exploitation and use.**

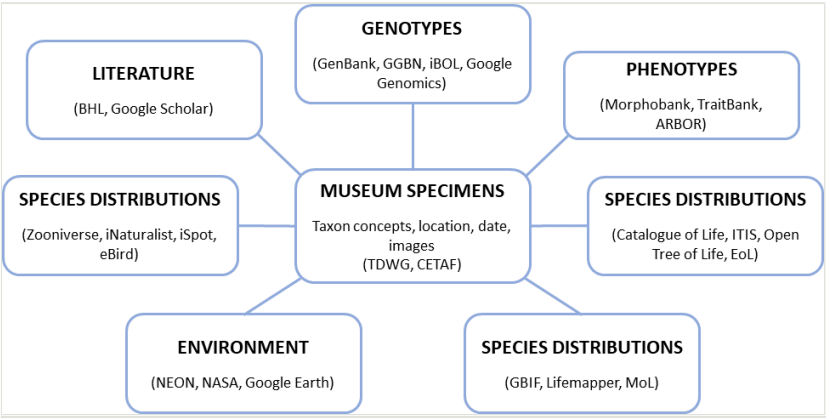

Figure 1. doi

NH informatics: Key links between museum specimens and other major data sources (Genotypes: GenBank, GGBN, iBOL, Google Genomics; Phenotypes: Morphobank, TraitBank, ARBOR; Species Distributions: Catalogue of Life, ITIS, Open Tree of Life, EoL; Species Distributions: GBIF, Lifemapper, MoL; Environment: NEON, NASA, Google Earth; Species Distributions: Zooniverse, iNaturalist, iSpot, eBird; Literature: BHL, Google Scholar).

SYNTHESYS+ will address these shortfalls, unlocking information in NH collections and linking it to other datasets to support innovative scientific research and to address major socio- economic challenges.

Specific Objectives

To accomplish the project aim of unifying collections operations and access, SYNTHESYS + will:

[**1] Invest in the development of balanced multi-modal access to collections to reach new and expanded user communities**, through delivery of an integrated programme of transnational (TA) and virtual (VA) access, incorporating users from new geographic areas as well as previously unrepresented research domains.(*measured by delivery of 7,017 days of TA prioritising new communities, as well as successful delivery of VA calls and tracking of outputs from both*)

[**2] Improve researchers' capacity to use collection information to tackle complex scientific challenges**, achieved by delivering a unified digital gateway to European collections, massively expanding the proportion of digital collections across Europe, and developing roadmaps for implementation of priority infrastructure components across Europe.(*measured by delivery of the European Loans and Visits System (ELViS), new digital content and an infrastructure roadmap white paper*)

[**3] Support the interplay of social and cultural aspects of collection data** by developing a pan European process for multi-stakeholder prioritisation of digitisation

activities and supporting the development of new categories of collections user.(*measured by delivery of a new multi-stakeholder forum and tracking of new research uses of collections*)

[**4] Bring scientific collections into the information age by investing in a linked open data approach** across our institutional data portals and implementation of common standards for molecular, digital and image collections.(*measured by adoption of these standards and processes across the consortium*)

[**5] Develop and implement targeted joint research agendas** supporting application of cutting edge artificial intelligence technologies to aid the mass mobilisation of collections data, new molecular processes to improve the speed and efficiency of generating molecular data from collections, and the development of high throughput 3D digitisation processes.(*measured by the number of specimens processed by these three new platforms and technologies*)

[**6] Identify collection data at European level and improve curation efficiency**, through developing collection data dashboards showing the contents of institutional collections and harmonising best practice in the provision of molecular and digital services.(*measured by delivery of the dashboards and efficiencies generated by adopting common molecular and digital standards across the consortium*)

[**7] Build and support new paths to industrial innovation** by incorporating a high proportion of SME's in the most technically innovative project components (ELViS and the Specimen Data Refinery) and engaging industry as users of biodiversity data infrastructure through the new multi-stakeholder forum.(*measured by the strength of SME partnerships and SME exploitation of collections data*)

[**8] Enhance digital skills and competencies, tooling-up researchers to navigate the big data domain** by delivering focused training activities.(*measured by the development and participation in the training programmes within and beyond the consortium and through the adoption of project platforms and services*)

[**9] Increase the engagement of society by providing alternative ways of benefiting from the national investments in collections and internationalising this effort**, to be delivered via foresight studies, building on the 2018 Global Biodiversity Information Outlook, engaging European and international experts, including data infrastructure such as EUDAT, EGI and the European Open Science Cloud.(*measured by the breadth and level of participation in the foresight studies and the delivery of white papers covering the interoperability and partnership potential for working with other infrastructures*)

By achieving these objectives, SYNTHESYS+, alongside related initiatives (ICEDIG; icedig.eu) will lay the groundwork for operational activities of the new integrated organisation DiSSCo (Distributed System of Scientific Collections, see letter of support) that is the European NH communities ESFRI initiative. **DiSSCo will take on the management and continued development of activities at the end of SYNTHESYS+**, supporting further technological and networking innovations necessary to efficiently provide

physical, digital and molecular access to 1.5 billion collection objects, and to consolidate the activities and governance of the participating organisations. When successful, the natural science community will be a fully enabled player in digital society, providing fundamental scientific data on the natural world, which will be freely and openly available for all.

Europe: a global leader

Natural science collections, which exist in all the world's countries, are some of the oldest RIs. Within Europe they include large institutions such as the Natural History Museum London and the Muséum national d'Histoire naturelle, both established in the 18th century, as well as many natural science museums, universities, botanical gardens and research centres, with their associated biological collections and research expertise. These institutions have always been open for all scientists and form the foundation of bio-and geo-diversity science that studies life on Earth, past and present. Initially they addressed fundamental questions in systematics, biogeography and conservation. While this remains a core mission, in recent decades European collections have taken on even greater significance (David and Taquet 2017). Many of them have turned their attention to **tackling the most important challenge that humans face over the next 30 years – the Anthropocene Challenge** –mapping a sustainable future for ourselves and the natural system on which we depend. This is a key moment for humanity; the global human population is predicted to peak in 2050, a fact that makes the next 30 years unique in the 200,000 year history of our species and the 10,000 year history of our civilisations. In this context, **NH collections are a key resource that can bring together human decisions for the short- to medium-term with an understanding of the mechanisms that determine the long-term impacts of environmental change.**

In support of this mission, European collections have a long history of working with associated e-infrastructures for accessing biodiversity data. For example, global organisations such as **GBIF**, as well as other related initiatives such as the **Catalogue of Life**, the **Biodiversity Heritage Library**, the **Encyclopedia of Life**, the **International Barcode of Life** and the **Biodiversity Information Standards organisation TDWG** all have European founding members from EU NH organisations and in many cases are headquartered in Europe. Collectively these organisations act as a loose consortium, each tackling their own special data types and services in partnership with collection-holding institutions. For instance, GBIF gives semantically harmonised access to over 37,000 databases with a total of >900 million biodiversity records, out of which about 150 million represent specimens from museum collections worldwide. **For the first time, the European branches of these global organisations have agreed to formally become part of the SYNTHESYS+ consortium, and will lead their respective interests and strengths within the SYNTHESYS+ work programme.** This is of critical importance, because there is currently a disconnect between these e-infrastructures and the physical collections infrastructure. **Only an integrated infrastructure, combining physical and digital access, is able to fulfil the long-term ambition of delivering sustainable and holistic access to European Natural Science collections.**

The low proportion of European museum specimens that are digitally accessible is a critical gap for achieving this vision of integration and accessibility. Museums and botanic gardens have keen interest in digitising their collections (Ang et al. 2013, Balke et al. 2013). The sheer magnitude of the task of digitising collections, however, is daunting. For example, with traditional methods, working one specimen at a time, one person can image and completely digitise the data associated with 50 specimens in a working day, with a basic cost of about 5€ per specimen (Koureas 2017) . It would thus take 100,000 person years to digitise one billion specimens,with a cost of 5€ billion. Organising such a task is beyond the capacity of individual institutions. What is needed is the transformation of a dispersed and fragmented access model to an integrated, data-driven RI that will bring the natural science collections into the information age, unifying access to the European collections and facilitating innovations that streamline digitisation of physical collections. This new RI project, the **Distributed System of Scientific Collections** (DiSSCo; dissco.eu), was accepted for inclusion in the 2018 ESFRI roadmap.

The user base

There is a wide range of traditional and new user groups for natural science collections. Conventionally, this community includes all researchers engaged in discovering, describing and interpreting life on Earth, both past and present, as well as researchers studying the geological diversity of the planet. According to recent survey data from the DiSSCo initiative (Casino et al. 2018), circa 16,000 researchers travel every year to physically access European natural scientific collections, and 800k objects are packed and shipped internationally (at an annual public cost of more than €70M). The majority of these specimens form part of life and earth science collections. For example, European collections hold circa 80% of the 2 million species presently described and **are at the forefront of efforts to describe what is estimated to be approximately 6 million new species that await discovery**. NH institutes also include extensive palaeontological and mineralogical collections, including concentrations of rock and ore samples, making them a valuable resource for the field of economic geology; exceptional meteorite collections used to study the origins of our solar system and how meteorite impacts could affect our future; ocean bottom deposits, critical for studies of the ocean and ocean floor, including research looking at global change, climatic warming and marine pollution; as well as extensive gem collections.

These extraordinary collections are physically distributed across institutions, but when taken together they present **an unparalleled source of scientific evidence about the natural environment**. They can be conceptually viewed along two fundamental time scales:

1. **Deep Time**: the dynamic history of change in the geology and life of our planet, spanning 4.56 billion years, including tectonic shifts in the continents, the rise and fall of major lineages of the tree of life, major shifts in species' geographic distributions, biomes, ecosystems and environmental signatures and trends.

2. **The Anthropocene**: the recent changes in biodiversity and ecosystem function resulting from the impact of modern humans on land use and resource exploitation, including species extinction, shifts in the distribution and abundance of species, climate change, and the emergence of new pests and diseases.

Collections are used to research a wide range of major scientific and socio-economic areas including:

- **Biodiversity discovery and conservation**
- **The genomic basis for the diversity of life**
- **Reconstructing the Tree of Life for all living and extinct species**
- **Understanding the co-evolution of Earth-life systems**
- **Modelling environmental, biotic and climatic change over the history of life**
- **Control of neglected and emerging tropical diseases and invasive species**
- **Ensuring sustainable agriculture and supply of raw materials**
- **Development of new diagnostic tools for biomedicine**

The unprecedented taxonomic, geographic, stratigraphic and historical coverage gathered together within these collections, coupled with their increasing digital accessibility, is **opening them up to entirely new user communities**. These users are increasingly drawn to the time series represented within these collections, to make predictions about the sustainable exploitation of bio- and geo-diversity that inform practical and policy decisions (Table 1),

Through innovative dissemination activities, SYNTHESYS+ will work closely with stakeholder organisations to promote the breadth of research uses for collections, highlighting their scientific and societal value.

1.2 Relation to the work programme

SYNTHESYS+ addresses the needs of the NH collections community in the environmental and earth sciencesdomain, as identified in topic INFRAIA-01-2018-2019: Integrating Activities for Advanced Communities. Ourworkplan will integrate and improve access to an expanded network of European and eligible third party NHcollections, and to their related instrumentation facilities, by:

- **expanding the user-base for NH collections data, by targeting access for new user groups such as those working in food and agriculture, medicine and health**;
- **developing innovative research services to meet the needs of a broader scientific community of users, targeting those addressing societal challenge topics in areas such as climate and food security**;
- **improving collections accessibility to a wider range of scientists, by expanding our TA programme and providing a new VA service**;
- **ensuring the long-term sustainability of these integrated services by transitioning them to the DiSSCo ESFRI infrastructure and aligning the work**

activities with the work-plans of leading global initiatives within the NH collections community.

SYNTHESYS+ builds on the work of prior SYNTHESYS projects that began to address the fragmentation of European collections. Through SYNTHESYS+ and its development under the preparatory phase of DiSSCo we will develop a unified and sustainable service that provides a one stop shop for European collections, collections data and the associated expertise.

An advanced community of key RIs

SYNTHESYS+ builds on strong foundations. The natural science collections in Europe have been working together for 20 years, initially through CETAF, and on a global scale as leading partners in a series of international initiatives, including Scientific Collections International (SciColl) with the mission to increase the use and impact of scientific collections for interdisciplinary research and societal benefits. The collective experience of partner institutions as sustained RIs spans several centuries. They directly employ more than 1,000 scientists, annually receive circa 16,000 scientific visitors (30-40% international), and have more than 100,000 scientific specimens on loan each year. As museums and botanical gardens, these institutes attract more than 10 million public visitors, inaddition to 25 million web visitors annually.

SYNTHESYS+ partners are home to cutting edge instrumentation, employing methodologies such as computer tomography, scanning electron microscopy, ultramicrotomy, X-ray fluorescence and X-ray diffraction, 3D imaging and mass spectrometry. Most include molecular biology laboratories and some include genomic service centres, often associated with partner locations that are operated by research consortia that facilitate interdisciplinary research and joint programming.

The natural science collections community has been described by the European Commission as a "super-advanced community", successfully implementing a series of I3 projects including three SYNTHESYS projects (2004-2018), to expand and enhance user access, as well undertaking a series of joint research activities. These projects have significantly enhanced the capacity of the consortium to interoperate, thus streamlining and enhancing user access. The larger community has also successfully initiated several EC FP supported projects focusing on scientific collections, including BioCASE, EDIT, EU BON, OpenUp!, PESI, ViBRANT and pro-iBiosphere.

Widening Transnational Access (TA)

The study of biodiversity is a highly cross-disciplinary activity. This requires the linkage of data across many fields from life and earth sciences, as well as the collaboration of scientists across broad domains of research (e.g. taxonomy, molecular biology, health science, industry, climate science, modelling, data science, programming and policy making). Prior instances of SYNTHESYS demonstrate an impressive track record of attracting diverse users to our collections, totalling more than 51,000 researcher days,

4,164 projects and approximately 4,000 research outputs to date (Fig. 2, Table 4). Recent examples funded from SYNTHESYS3 include projects to understand the distribution and transmission of West Nile Virus from mosquito collections (AT-TAF-3844); indicators of plastic and metal pollution by studying the dental and skeletal pathology of marine mammal collections (DK-TAF-5825); archaeological studies on the use of flax products for textile and oil production in ancient societies (DE-TAF-3837); studies on 16th century Spanish and Portuguese botanical illustrations to identify cultural links between the two countries (ES-TAF-3724); detection of contaminant seeds in Sardinian irrigated crops to decrease the impact of alien plant species (GB-TAF-4276); and the detection and timing of Iron oxide-copper-gold type mineralization in the SW East European craton to assess mining and mineral potential (SE-TAF-7040).

Table 4.

Applicant nationality data from SYNTHESYS3

Country	Eligible - Call 1	Accepted - Call 1	Eligible - Call 2	Accepted - Call 2	Eligible - Call 3	Accepted - Call 3	Eligible - Call 4	Accepted - Call 4	Total Eligible	Total Accepted	Percentage of eligible applications accepted
Spain	84	30	91	30	87	33	85	39	347	132	38%
Italy	77	32	82	34	80	38	92	33	331	137	41%
Poland	78	32	73	20	69	29	75	34	295	115	39%
Germany	50	27	52	24	38	17	53	31	193	99	51%
Others	47	14	52	18	39	16	44	19	182	67	37%
France	40	20	39	19	31	15	40	22	150	76	51%
United Kingdom	33	16	33	15	35	19	44	22	145	72	50%
Portugal	48	17	39	11	19	5	28	8	134	41	31%
Czech Republic	38	21	31	14	37	19	24	12	130	66	51%
Bulgaria	27	9	30	13	22	8	20	9	99	39	39%
Hungary	33	9	29	7	14	2	21	9	97	27	28%
Turkey	14	2	24	3	26	6	14	4	78	15	19%
Netherlands	25	11	16	6	19	7	13	8	73	32	44%
Greece	11	3	11	4	20	8	16	7	58	22	38%
Republic of Serbia	13	5	18	7	12	5	12	5	55	22	40%
Slovakia	11	3	12	2	15	5	13	3	51	13	25%

Country	Eligible - Call 1	Accepted - Call 1	Eligible - Call 2	Accepted - Call 2	Eligible - Call 3	Accepted - Call 3	Eligible - Call 4	Accepted - Call 4	Total Eligible	Total Accepted	Percentage of eligible applications accepted
Romania	14	1	15	4	7	2	14	4	50	11	22%
Croatia	12	5	11	4	10	2	10	6	43	17	40%
Belgium	8	4	17	7	6	2	10	8	41	21	51%
Austria	7	3	9	5	16	3	7	1	39	12	31%
Denmark	5	3	6	2	10	2	11	3	32	10	31%
Israel	7	2	8	3	7	2	8	7	30	14	47%
Norway	2	0	4	2	5	2	10	5	21	9	43%
Switzerland	4	1	7	2	5	2	3	2	19	7	37%
Sweden	4	0	4	1	4	3	7	1	19	5	26%
Lithuania	2	1	5	2	3	1	5	2	15	6	40%
Finland	4	1	4	1	2	1	1	0	11	3	27%
Ireland	1	1	1	0	4	1	4	1	10	3	30%
Macedonia	3	3	3	1	3	2	1	1	10	7	70%
Latvia	0	0	3	0	4	2	2	0	9	2	22%
Estonia	1	0	4	2	3	3	0	0	8	5	63%
Albania	2	1	1	0	1	0	1	0	5	1	20%
Iceland	3	1	0	0	1	0	1	0	5	1	20%
Slovenia	2	1	2	1	0	0	1	1	5	3	60%
Cyprus	0	0	1	0	1	0	2	0	4	0	0%
Bosnia and Herzegovina	0	0	2	2	1	0	0	0	3	2	67%
Liechtenstein	1	0	1	1	0	0	0	0	2	1	50%
Faroe Islands	1	0	0	0	0	0	0	0	1	0	0%
International Organisation	0	0	0	0	0	0	1	0	1	0	0%
Luxembourg	0	0	0	0	0	0	1	1	1	1	100%
Montenegro	0	0	0	0	0	0	1	0	1	0	0%
Malta	0	0	1	0	0	0	0	0	1	0	0%

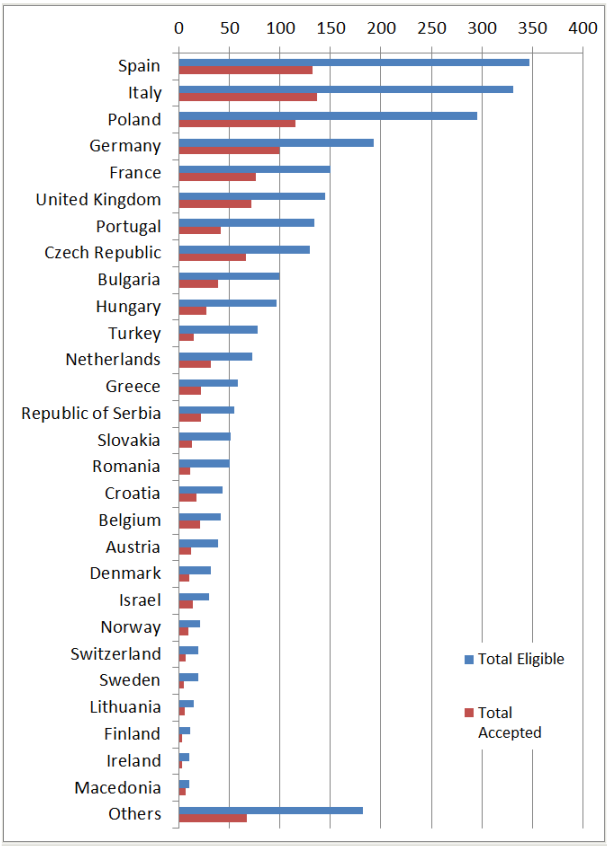

Figure 2. doi

SYNTHESYS3 applicant nationality for Transnational Access Calls (2012 – 2016). 43% of successful applicants were female, against an 33% average for scientific researchers (EC Directorate-General for Research and Innovation 2016).

In SYNTHESYS+ we are expanding the network of partner institutions by targeting those with new collections that provide unique geographic, taxonomic and stratigraphic representation. This expanded coverage provides a single point of entry to the largest and most comprehensive network of collections in the world. As part of this effort, and new for SYNTHESYS+, we will have a special focus on new user communities, widening access to the collection through targeted interventions to strengthen and broaden our user base. Approaches will include:

- Up-weighting the evaluation score of access request proposals **from new user communities**, to balance opportunities for these groups against those from experienced, well established domains;
- Up-weighting access applications focusing on societal challenge topics, such as environmental & climate change, biodiversity planning / monitoring, food security & agriculture, forestry, and health;

- New advertising and dissemination activities at key conferences, promoting SYNTHESYS+ as a single point of entry to European collections;
- Targeted promotion for new research communities, especially those in societal challenge areas, illustrating the new paradigms of research (biodiversity modelling, citizen science, measuring hyper diversity, automated species recognition) made possible by new technologies exploiting NH collections;
- Live public data / dashboarding of SYNTHESYS+ research projects and activities, illustrating the tremendous diversity with which NH collections are being exploited by diverse user communities, and providing monitoring to ensure use is balanced across stakeholder groups.

Through the SYNTHESYS+ public dashboard, the senior management team will monitor, and where necessary manage, access requests to ensure the network promotes equal opportunities across the Access programme, as discussed in our gender and diversity action plan (see Management work package description). Likewise, the management team will maintain the evidence and documentation necessary to demonstrate the impact of this plan.

New Virtual Access (VA), supporting Digitisation on Demand

For the first time, building on preparatory research in SYNTHESYS3, SYNTHESYS+ will offer an integrated Digitisation on Demand (DoD) service, creating an entirely new service class that provides virtual collections, freely available to a global user community. An independent committee will be used to prioritise requests for virtual collections access through the associated Joint Research Activity (task 7.1). This VA programme will then fulfil prioritised requests, undertaking the necessary digitisation, data curation and provisioning to make the data accessible through open public portals. This activity will make use of an 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. For example, NHM London alone has 9.1 million records available, which have been downloaded more than 10.1 billion times since April 2015, delivering more than 60 tracked publications. The primary constraint on the exploitation of these portals is the data they contain, as this usually represents a small fraction (4.5-18% for the eight largest collections) of a given institution's physical holdings. **VA within SYNTHESYS+ will address this critical gap, helping institutions to increase their proportion of digital collections in response to evidence of a strong demand by current, emerging and new user communities, and to ensure free access to all users.**

As part of this activity, all VA requests will be publicly tracked and monitored through a data dashboard, with metadata (minimally, number of specimens digitised, number downloaded, data cited and number of publications) submitted to an external board, for assurance and to provide value for money recommendations to ensure institutions and the research users make the best possible use of the service.

Provision of VA will exploit the existing digitisation capital equipment within participating institutions, **with EU financial support only being used to cover the technological and scientific support activities needed by researchers to effectively use this service.**

Harmonisation, standardisation and internationalisation of operating activities

Working towards ultimate operation under the umbrella of the DiSSCo ESFRI initiative, **SYNTHESYS+ networking activities are primarily geared toward increasing levels of harmonisation and standardisation in data and services**. This will increase the efficiency of partner institutions, which in many cases continue with artisanal methods of operation due to local communities of practice, or to differences in national and regional policy. Increasing standardisation also facilitates better integration of SYNTHESYS+ products and allows us to work toward standard operating procedures across our institutions, which can be integrated with training packages to spread best practice.

SYNTHESYS+ partners already have a very strong culture of cooperation, including associated infrastructures, scientific user communities and industry stakeholders. **The SYNTHESYS+ networking activities have been designed to strengthen these relationships and build upon them as a precursor to operation under DiSSCo.**

Within SYNTHESYS+, this will be led by the major European and international organisations (CETAF, GGBN, TDWG and GBIF) covering their respective areas of interest (European collections, genomic biobanking, digital standards and biodiversity data sharing). These organisations are highly experienced and internationally recognised as world leaders within their respective domains. Embedding SYNTHESYS+ activities into the strategies of these organisations increases the sustainability of the overall programme of work as well as supporting greater community integration towards the ultimate goals of DiSSCo. These organisations also provide linkage to communities outside Europe, providing new opportunities for dissemination and promotion of SYNTHESYS+ goals and vision.

Networking and integration activities within SYNTHESYS+ are geared toward the adoption and deployment of global standards, particularly in the areas of data mobilisation, curation, preservation and provision, ultimately aimed at improving access to data produced through the project. As in SYNTHESYS3, data generated will be in line with FAIR principles and “open by default”, with exceptions only granted in specific cases such as exclusions for personal identifiable information or other ethical considerations. Data will be managed and tracked through data management plans, which will be a condition for all users requesting access. Networking activities include a programme of dissemination and promotion, particularly focused on innovations developed through joint researchactivities such as molecular and digital workflows, and by reinforcing innovations developed with our industrial partners. This includes promoting the benefits of the centralised European Loans and Visits System (ELViS) (JRA1) and Artificial Intelligence (AI) activities, associated with the Specimen Data Refinery (JRA3). Our management plan includes special activities to increase the representation of women and younger scientists within the consortium and amongst SYNTHESYS+ users. This especially applies to areas such as AI, where the

community needs to attract a pool of students to develop their computer science skills against our use cases.

Knowledge transfer activities, especially related to internationalisation of SYNTHESYS+ and the development of help desk/training packages, are scheduled as part of the Distributed European School of Taxonomy. This will deliver training on the latest molecular and digital standards and protocols developed through SYNTHESYS+, especially those aimed at improving interoperability of data within the molecular (NA3) and digital (NA4) work packages. Networking activities include registration of digitisation, data processing and biobanking services in European service catalogues, including the European Open Science Cloud (EOSC). In addition to aiding service discovery, this work will provide the means to benchmark SYNTHESYS+ activities against similar international services. Foresight studies, building on the 2018 Global Biodiversity Information Outlook (GBIO), will be conducted to enhance and expand the GBIO roadmap in sectors relevant to the NH collections community, strengthening the deployment of SYNTHESYS+ services and developing a clear plan to sustain and extend this work on project completion. An additional goal from networking activities is to embed SYNTHESYS+ activities within the DiSSCo programme, ensuring a smooth transition to DiSSCo operation.

Innovating and improving services

SYNTHESYS+ Joint Research Activities will specifically target gaps in current service provision. This includes the development and refinement of high throughput processes, such as pipelines for 3D models of NH specimens, and protocol infrastructures to support molecular sequencing-on-demand. While high quality services in these areas have been operational across many SYNTHESYS+ partner institutions for over a decade, they do not operate at a scale or with the degree of integration necessary to generate the benefits envisaged by DiSSCo. **By developing standard operating procedures across institutions, we will be able to dramatically increase the rate of sample processing, while retaining high quality results, and benefit from economies of scale generated through common practices**. In development of these services, SYNTHESYS partners will work closely with complementary research and design studies (e.g. ICEDIG), as well as other projects in the cultural and ICT sectors undertaking related activities.

The development of ELViS is central to achieving the vision of integrating all forms of access to collections. ELViS is intended to become the gateway to European collections and will be co-developed with a commercial partner (Picturae), founded on a functional expansion of the existing platform used to manage TA requests in SYNTHESYS1-3. The platform will be completely redeveloped, using agile software development practices. All collection-holding partners will be involved in specification and testing phases, with core functionality to support physical access requests, and Collections on Demand prioritisation from year 2 of the project. Ultimately it is envisaged that ELViS will be expanded to integrate with institutional collection management systems, but this advanced functionality is expected under DiSSCo. In support of this vision, the ELViS API will be developed and documented in the latter stages of SYNTHESYS+.

A key limiting factor in organising and using information from global NH specimens is making that information computable. More than 95% of available information currently resides on labels attached to specimens or in physical registers. Institutional digitisation pipelines have tended to focus more on the specimens themselves than on efficiently capturing computable data about them. **SYNTHESYS+ will address this gap using technologies developed to harvest, organise, analyse and enhance information from other sources (such as books, photographs and maps), offering the prospect of greatly accelerated data capture**. Technologies of particular interest include computer vision, optical character recognition, handwriting recognition and language translation. Prototyped through SYNTHESYS3 activities, application of **these technologies will now be developed into a cloud-based platform (the Specimen Data Refinery), that will bring together workflows for processing specimen images (particularly label images) by semantically tagging image components and then submitting them to appropriate additional processes for data extraction**. Working with commercial partners in the field of handwriting and optical character recognition, we anticipate piloting the Specimen Data Refinery with several hundred thousand specimen and label images, **in anticipation that this will become the pivotal technology to achieve rates of digitisation expected through DiSSCo**.

Relationship with related ESFRIs and ICT / e-infrastructures

The European landscape of RIs is changing rapidly. Continuous investments at national and European level, in conjunction with a common vision to deliver more impactful and efficient tools supporting research and innovation, are gradually transforming a fragmented landscape into a more coherent space where scientists can access services as part of the European Open Science Cloud (EOSC). For environmental RIs to provide high quality services, it is imperative to semantically link primary collections data (e.g. taxonomic and trait datasets) with ecological, geological and climate datasets. Mobilising and publishing primary information is therefore a crucial step in supporting the operation of many other environmental RIs, not just NH collections.

The Environmental cluster of RIs (ENVRI) has also identified the importance and position of primary biodiversity information in supporting the operations of other Environmental RIs (Fig. 3).

SYNTHESYS+ will leverage the TA and VA programme by putting in place mechanisms (tools and workflows), which will improve the FAIRness of data produced by the study of collections. In this way, the SYNTHESYS+ consortium will spread good practices among collection users and collections holders. These best practices will subsequently support the development of the pan-European RI for collections (DiSSCo). SYNTHESYS+ will significantly advance the way that the NH community interacts with the European Open Science Cloud. By developing service components that can be directly incorporated in the EOSC marketplace, SYNTHESYS+ will work together with core e-infrastructures (i.e. EGI nodes) to develop common authentication and authorisation mechanisms for all anticipated e-services, such as those required by ELViS. Finally, SYNTHESYS+ will deliver its

networking work programme in close collaboration with key international infrastructures and networks, ensuring optimum alignment. For example, the internationalisation activities (NA5) will be led by GBIF. **This will ensure that European level activities are synchronised with global initiatives.**

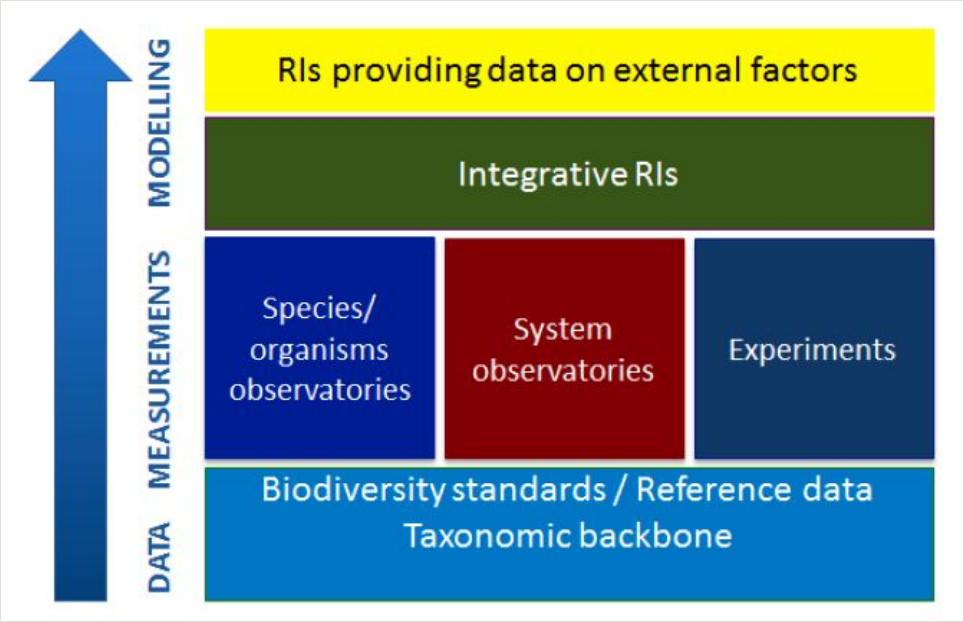

Figure 3. doi

Biodiversity information in the context of RI operations (data-modelling) by the ENVRI community.

Partnership with industry

Interest in and use cases for SYNTHESYS+ from industry / private sector stakeholders are manifold, currently representing approximately 10% (circa 50,000 users) of the user base for European collections (Koureas, 2018). Partnerships include service providers; industrial device/technology producers; and sectors including health, mining and others. Potential practices range from the use of genetic resources and feasibility of linking genes and phenotypes, to the development of technologies and materials, the use of images and videos for marketing, and the link between natural history and sustainable tourism. **Applied research with industry is leading to innovation in areas such as agriculture, forest management and environmental assessment, directly contributing to tackling societal challenges.**

The development and installation of ELViS, the key software supporting SYNTHESYS+ users, is being provided by an industrial supplier (Picturae). In addition, specialist SME partners will be providing digitisation technologies, genomics skills, ICT support and software development. Industrial partners in SYNTHESYS+ will represent the domains of

artificial intelligence and machine learning, as well as optical character recognition and natural handwriting recognition. Working in close partnership with these companies ensures that the very specific needs of SYNTHESYS+ can be fulfilled by SMEs, who specialise in niche market innovation and the provision of tailored solutions. Industrial partner activities during the project will support knowledge transfer and other dissemination activities (e.g. collaboration at trade fairs and participation in advisory bodies) to build lasting relationships. For example, we will include industrial stakeholders on the board of the external committee overseeing prioritisation of DoD requests, to ensure that the needs and interests of industry can be met. This deeper incorporation of industrial partners within the SYNTHESYS+ programme is a first for SYNTHESYS, and a critical transition towards ever-greater industrial cooperation within DiSSCo.

Measuring progress and independent assessment

SYNTHESYS+ will use a combination of quantitative and qualitative measures of success, spanning the way in which partners operate, how data is used, research innovations, and public impact, including:

- Transformation of collection holding partners with respect to data sharing (e.g. number of participating institutions sharing data, volume of uploaded data and metadata, data sharing across institutions);
- Demonstrated utility of tools and services that link data and catalyse scientific research (e.g. number of publications, software downloads, deployment across and beyond the consortium, API calls);
- Rate of digitisation of NH information (i.e. number of specimens digitised, rates and costs of digitisation per specimen), through the Specimen Data Refinery;
- Total data available through the ELViS platform (e.g. number of specimens with core metadata, number of specimens with genomic and / or 3D datasets);
- Research and public exploitation of data via the ELViS platform (e.g. data downloads and citations, links from other data repositories) including the number of research outputs tracked by ELViS (e.g. research papers, collaborations and grant applications);
- Novel public-driven use of data from the platform, including human-in-the-loop activities associated with the Specimen Data Refinery, as well as citizen science applications;
- Crowdsourcing transcription activities, and community-driven social media campaigns;
- Research breakthroughs (e.g. major papers in leading multidisciplinary journals, impacts on policy);
- Successful handover of platform to DiSSCo, together with a robust business model to ensure long-term support and development.

Through activities in NA2, SYNTHESYS+ will develop a series of public dashboards containing monthly updates to these key metrics, shown as data visualisations, to highlight long-term trends. This methodology was pioneered in SYNTHESYS3 and exploits

commercially available business intelligence software that is widely used by several SYNTHESYS+ partners.

The vast scale of our collections and the high levels of demand from our diverse user community mean we need to develop mechanisms to triage and prioritise requests to support Collections on Demand services, such as sequencing, digitisation and transcription. To facilitate this, SYNTHESYS will bring together panels of experts, including representatives from industry, who will agree and implement a set of criteria for prioritising VA and TA requests. Panels in each domain will assess the quality and feasibility of the proposed research outcomes. For VA, the institutional benefits of mobilising the data will also be considered. After the first year, the ELViS platform will take on the role of maintaining documentation to support and justify access requests. This will include records of the names, nationalities and home institutions of applicants within the research teams, as well as the nature and quantity of access provided to them by their host institution. In the interim, the online system used in SYNTHESYS2 and 3 will fulfil this role for the first Access call.

Progress beyond current achievements

In summary, **SYNTHESYS+ represents a major enhancement over prior iterations of SYNTHESYS, that will be a step-change in providing a critical pathway toward preparing the collections community for the DiSSCo RI**. The innovative new VA programme and associated digitisation prioritisation activities will be **a world first, providing the means by which any user can request digital access across multiple European institutions**. This will create new paradigms of research opportunity that are expected to have a transformative effect on our institutions and users. The expanded TA programme, promoting involvement of new user communities, will widen access as never before, providing physical access opportunities when VA is not scientifically useful or technically feasible.

SYNTHESYS+ **innovation activities address critical gaps in our technical infrastructure**, such as development of ELViS as a common gateway to European Collections. Developments such as the Specimen Data Refinery, alongside improvements to digital and molecular workflows, will embrace the latest technologies with commercial partners, to support the rates of specimen processing necessary to transform European NH collections. Our networking activities, led for the first time by the key networking consortia in each domain relevant to NH collections, will ensure that the SYNTHESYS agenda is promulgated and developed as a common enterprise beyond Europe. Finally, the embedding of SYNTHESYS+ as a critical delivery project supporting DiSSCo will enable a step-change in the level of integrated governance and security, ensuring SYNTHESYS activities can be sustained for the long term, and that the transformative vision for DiSSCo can be achieved. **In sum, these novelties make SYNTHESYS+ the largest, most ambitious and arguably the most important iteration of SYNTHESYS since its inception in 2004.**

1.3 Concept and methodology

(a) Concept

A recent survey of European collection-holding institutes (Koureas, 2018) revealed that two-thirds of all collections (89 institutes), received a total of 16,500 visiting researchers annually and loaned more than 760,000 specimens each year. These activities cost circa €11.5 million annually, alongside infrastructure and software investments of approximately €1.75m per year. Costs for digitisation (€11m), information management software (€12m) and digital curation (€5m) give a total annual expenditure of about €52m for these 89 institutes. Despite these investments, the current fragmented model for managing collections in Europe limits researchers in discovering and accessing specimens, limits economies of scale and scope for research in collections and is a major obstacle for providing balanced multi-modal access to collections in Europe. For example, the €12m in annual software expenditure is caused by the wide variety in collection management systems deployed at these institutions. Given this, **a pan-European RI uniting natural science collections is urgently needed to overcome the current limitations on the use of collection data and related expertise for societal benefit.**

SYNTHESYS has made enormous progress toward this goal over the past 14 years, in terms of defragmenting activities, integrating access and using joint research activities to innovate, filling critical gaps in our infrastructure. But even among SYNTHESYS partners, institutional governance ultimately rests with participating institutions, and constraints on the practical size of the consortium mean that SYNTHESYS+ only covers a fraction (21) of the estimated 140 collection-holding institutions across Europe. The DiSSCo pan-European research ESFRI initiative offers the prospect of bridging this gap - centralising institutional governance and providing greater coordination, allowing all collections to be viewed as a single common enterprise rather than the assets of a series of cooperating institutions. With this in mind, **SYNTHESYS+ has been developed as a critical precursor to preparing the major partner institutions for the transition to DiSSCo.**

An introduction to DiSSCo and its relationship to SYNTHESYS+

The DiSSCo project was accepted onto the 2018 ESFRI roadmap and builds on a mature community of institutions, including all SYNTHESYS collection-holding partners. It is a collaboration of 115 national facilities in 21 countries (Fig. 4), the largest ever formal agreement between natural science collection facilities. This strategic collaboration is underpinned by the sound governance and decision-making structures founded within SYNTHESYS. Tasks within SYNTHESYS+ are aimed at increasing cooperation and standardisation of practice to support the centralised governance model envisaged for DiSSCo.

The DiSSCo RI will unify access to information, to provide massive new linked data associated with collections, and to drive policy and process harmonisation. This new RI achieves the economies of scope and scale necessary to maximise impact for science and

society. Through digitisation, aggregation and the linking of European collections, it will be possible to draw critical new insights, enabling scientists to address some of the world's greatest challenges on a scale above and beyond what has been possible through SYNTHESYS. Critically, DiSSCo brings a transfer of authority from facilities to a central hub for all key operations, a clear decision-making mandate, and binding institutional commitments as part of a new independent legal framework. While SYNTHESYS has been working toward these goals, it has taken the proposal of a more inclusive infrastructure and changes in institutional governance to make DiSSCo possible. Having developed this agreement through 2015-16, the DISSCo proposal was formally announced on the ESFRI roadmap in September 2018.

Figure 4. doi
DiSSCo consortium

DiSSCo is uniquely placed among other distributed networks or infrastructures, filling a significant gap in data quantity and quality in the landscape of European RIs (Fig. 5). For example, LifeWatch is an ESFRI infrastructure that has a MoU with the DiSSCo initiative. DiSSCo and LifeWatch are complementary, as LifeWatch covers both modern observations that are born digital and DiSSCo historical data that still needs to be processed. Both feed into the global e-infrastructure of GBIF and provide services to GBIF users.

The DiSSCo preparatory and construction phases are planned between 2018 and 2025. In the initial phase of the programme DiSSCo will deploy systems and processes which have already achieved a high Technology Readiness Level (TRL; European Commission 2017), such as high throughput herbarium sheet digitisation which is very mature (TRL 7-8). However, for many areas (e.g. insect digitisation, VA and DNAoD), TRLs are much lower. In these cases SYNTHESYS+ will support the additional innovation, pilots and testing

programme necessary to raise the technology readiness for wider deployment across the consortium. **Thus, SYNTHESYS+ is a critical part of the innovation and testing programme for DiSSCo, providing an expanded and well tested evidence-base for DiSSCo's development and operation.**

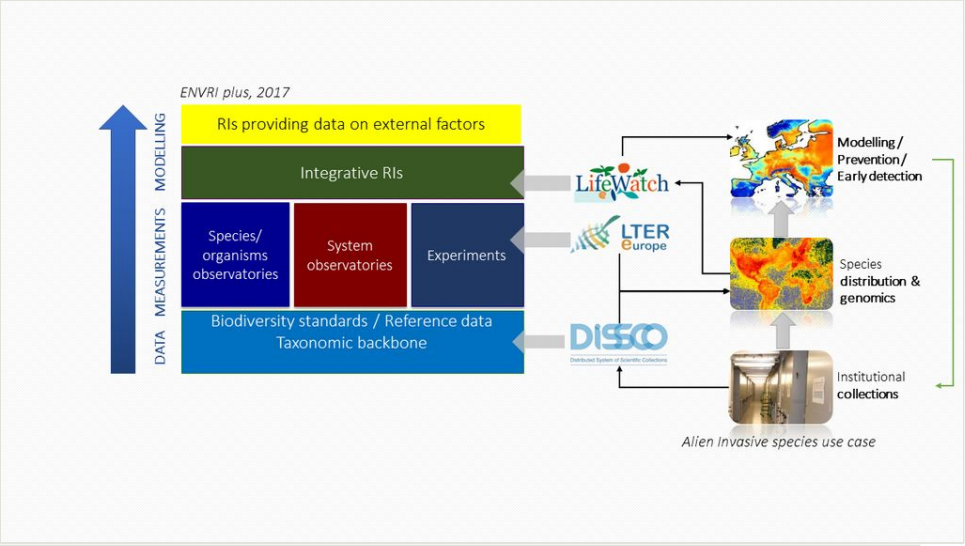

Figure 5. doi

DiSSCo service relationships with other research infrastructures

(b) Methodology

The SYNTHESYS+ programme of work comprises a series of interlinked activities, which build on the work of SYNTHESYS1-3 and fill critical gaps in the development of the DiSSCo RI. Activities have been carefully designed to form a coherent and coordinated set of components to deliver the SYNTHESYS+ objectives. To help navigate this complexity, activities have been structured into three streams, representing the three RI activities: NA, JRA and Access. Each will have a stream leader whose role is to oversee the work and interlink activities with the work of other streams. This is summarised in Table 2 and then explained in more detail throughout Section 1.3b.

1.3.1. Networking Stream

NA2 - Harmonisation of policies, best practices, training and support

NA2 comprises networking actions addressing the policies, best practices and standards that our scientific community needs to be aware of, adopt, and implement at institutional level. This will constitute a thorough, harmonised knowledge base on which further collaborative developments can be built and will be complemented by **integrating the four collections assessment methodologies which are currently deployed by NH**

collections worldwide. This will provide the first, integrated overview of global collections and will be developed in close association with JRA1 for integration with the ELViS platform for European collections. Dashboard data will ensure that potential access users have a clearer picture of the complete collections holdings within each SYNTHESYS+ institution, rather than just the small proportion currently digitised.

Training will provide the best tools, procedures and systems to enhance daily practice, integrated into an internationally shared framework. This will be delivered through a series of expert workshops and sessions, bringing together the experiences of key parties to identify priorities and key topics. NA2 will use the Teamwork software platform (part of the DiSSCo infrastructure currently used to manage projects) in support of consortium communication and dissemination activities.

NA3 - Molecular standards and processes

The Global Genome Biodiversity Network (GGBN) aims to foster collaboration among genetic repositories (biobanks) in order to comply with standards, best practices, interoperability and exchange of material in accordance with national and international legislation and conventions. The majority of GGBN members are NH collections, and include 15 partners in the SYNTHESYS+ consortium. Other members include culture collections, zoos and other repositories. Developing a network of trusted collections, establishing standards, and identifying best practices by reaching out to other communities are key aspects of GGBN's mission. This is especially critical in the light of new international legislation such as the recent Nagoya Protocol on Access and Benefit Sharing (ABS). The work planned for NA3 therefore benefits both SYNTHESYS+ and the wider GGBN.

NH collections are facing a series of challenges, including the rapid acceleration in sequencing technology. This has added access demands on parts of the collections that even just a few years earlier were considered unsuitable for genetic use. ABS legislation applies to nearly all collection types, and with biobanks increasing in number worldwide, **there is an urgent need to streamline procedures and to ensure legislative compliance**. Within Europe it is necessary to **1) reach agreement on common standards for biodiversity and environmental biobanks; 2) define best practices for the use of molecular collections; and 3) try to ease exchange of samples and related information while staying compliant with legislation and conventions**. NA3 will address these challenges by reaching out to other communities for knowledge exchange. After a landscape analysis involving all NA3 partners and international stakeholders, workshops and handbooks will be developed to collate existing standards and best practices. This will be used to develop a best practice framework. While NA3 is focusing on molecular collections, NA3 activities feed into i) NA5 towards a better international biodiversity infrastructure; ii) NA4 to find gaps in data exchange standards for molecular data; and iii) JRA2 to provide the curatorial basis and legislative compliance, while conducting sequencing on demand tasks. These linkages, coupled with the teaming up of GGBN, CETAF and GBIF as part of SYNTHESYS+ offers the prospect of achieving a high

level of international adoption and sustainability for the best practice standards and protocols developed though NA3.

NA4 - Digital standards and processes

Convergence on standards and processes is essential if digital records are to bridge across physical collections to achieve their full scientific potential. These standards facilitate seamless interoperability between collections and allow much more rapid and repeatable access to data. **NA4 will bring SYNTHESYS+ organisations together under the guidance of TDWG**. We will build on existing standards for collection management, but also draw on other fields that use collections data, for example invasion biology, conservation ecology and climate research. Some potential standards exist as frameworks in distinct domains and now need development to make them more generally applicable, while other standards need developing de novo, because they are novel applications of data.

An example of novelty is in the semantic linking of collections through the stable identifier framework (NA4.2) (Güntsch et al. 2017; Groom et al. 2017). Current implementations are similar to one another, because institutions look to each other for leadership on formats and controlled vocabularies. As more institutions deploy these identifiers across their institutional data portals, the community has indicated a need for more structure and coordination than is possible through an informal network (Guralnick et al. 2015). This includes minimal core standards to ensure compliance, monitoring to alert institutions when standards are compromised, and options to expand beyond core metadata elements. Development of this enhanced framework in SYNTHESYS+ will be through international meetings and development of pilot projects with testing using real data from diverse stakeholders. Processes and standards will be generated through an iterative process of discussion, testing and refinement until a proven workable consensus is reached that fulfils the goals of each task.

Other standards and processes are already more mature. The adoption of these will require a more educative approach with partners coming together to learn from each other and experimenting with the system using their own data and infrastructure. This will be underpinned by training activities in NA2, especially in areas such as VA support and for 3D digitisation (JRA2).

NA5 - Internationalisation and New User Communities in EC Priority Areas

SYNTHESYS+ needs to operate as an effective component in a much wider landscape of global investments. There are significant opportunities to increase efficiency and sustainability via collaborations with major international communities and infrastructures. **NA5 will deliver this connectivity by building on the Global Biodiversity Informatics Outlook (GBIO) report (Hobern et al. 2012) and recommendations of the 2018 GBIC2 workshop (Hobern et al. 2019). These landscape reviews identify a series of major component areas requiring coordinated planning and activity, and spanning aspects of research culture, data management, and analysis and use**. SYNTHESYS+ is well

placed to follow up and expand on aspects of this framework through support for the international coordination mechanism of GBIO, and deliver a series of detailed roadmaps for component areas. **These roadmaps will cover five to ten year periods and identify investments and developments considered necessary or beneficial to the global community.**

Workshops will combine expertise from within SYNTHESYS+ and other European stakeholders, as well as international initiatives (GBIF, Catalogue of Life, Biodiversity Heritage Library, International Barcode of Life), alongside other relevant international institutions and programmes. Across each continent, major national and regional communities have been undertaking activities comparable to those of the SYNTHESYS programme. For example, the NSF funded ADBC Programme, and in particular the iDigBio hub in Florida, undertakes work very similar to SYNTHESYS, while national biodiversity informatics initiatives such as ALA in Australia, CONABIO in Mexico and SANBI in South Africa, seek to deliver many of the same products and services. In coordination with CETAF (NA2), GGBN (NA3) TDWG (NA4) and DiSSCo, NA5 will develop a stakeholder forum to align outputs and share best practice across these continental pillars of activity. NA5 will also review opportunities for alignment, cooperation and increased sustainability between major European infrastructure stakeholders, such as DiSSCo, LifeWatch, EMBRC, ELIXIR, e-RIHS and eLTER. This approach will increase coordination and collaboration between SYNTHESYS+ partners and external communities, improving the efficiency, relevance and development of RIs supporting European collections.

NA5 will operate by supporting a Biodiversity Informatics Planning Office with GBIF to propose required components and practices around major areas of collections-focused informatics and through a series of workshops and foresight studies, to deliver a roadmap for each area. This office will:

1. Set up a series of international workshops to evaluate needs and issues and to propose required components or practices around major informatics areas (probably at the scale of the GBIO components);
2. Coordinate post-workshop consultation and refinement of proposals to deliver a shared international vision and agreed set of implementation needs as a roadmap for each component area;
3. Provide coordination and support for any parties preparing proposals and seeking funding to implement any of these needs;
4. Provide services to projects and infrastructures on alignment with the shared international vision and individual roadmaps.

Each workshop will convene a range of SYNTHESYS+ partners relevant to its topic and will fund the engagement of a selected set of attendees jointly to develop the resulting roadmap recommendations document. This work will deliver an integrated white paper outlining opportunities for European leadership in delivering key components within the roadmap plans.

1.3.2. Joint Research Activities Stream

JRA1 - Optimisation of Access

To address the challenge of optimising access as set out in section 1.3a, **an integrated, data-driven RI is urgently needed to bring natural science collections into the information age**. JRA1 will provide the first key component for this with the development of ELViS, the European Loans and Visits System. Experiences with development of a simple unified system for TA in SYNTHESYS2 and 3, coupled with the French development of Colhelper, the loans and visits system for one of the world largest collections (MNHN), demonstrate that the development of ELViS is feasible. Although development of a pan-European system is a major challenge, given the diversity of the collection facilities and the cultural change required by users, the costs for development are relatively low (annual €250K Euro for four years) and potentially generate significant savings across the community. The system is intended to gradually replace institutional loans and visits management systems during the preparatory phase of DiSSCo. Its modular, agile, microservice oriented development will provide the flexibility and sustainability necessary to promote adoption. **ELViS will be developed to be compliant with the new EU Data Protection Regulation, which will lead to future compliance with data protection regulations across all European collection facilities.**

ELViS has been planned to provide a one-stop shop for researchers to access physical collections as well as to support DoD requests by SYNTHESYS+ users. Sustainability will be assured by transferring maintenance and development to the DiSSCo consortium after the project, providing a future development path and adoption by the wider network of DiSSCo partners. ELViS will be developed by a commercial partner (Picturae), helping to ensure a mature, high quality open source software product is delivered that meets community needs. Picturae have been involved in many large-scale digitisation and software development projects across the globe, including several projects for NH collections. Their experience minimises the risk profile for ELViS development and will help to assure the community of their capacity and commitment to build the software. The system will include a ticketing system for user support and general service management; a help desk to support VA and TA work packages; and a dashboard to monitor VA and TA services, showing institutional collection capabilities and providing collection quality indicators. The help desk and dashboard will be operated by and developed in close collaboration with NA2. **The system will be designed with the capacity to operate multilingually to support international access requests from researchers worldwide.**

A pilot is included in JRA1 to study how to leverage the federated authentication and authorisation infrastructure (AAI) developed as part of the EOSC-hub, into ELViS. This will provide for future integration with national infrastructures and development towards a common access system for all Environmental RIs. Authorisation will be required to gain access to sensitive collection information and for the AAI pilot, ELViS services will be designed based on the blueprint developed in the Authentication and Authorisation for Research and Collaboration (AARC) project. The pilot will implement a cross-infrastructure case study to show transparent AAI interoperability between ELViS as a DiSSCo RI

service, and EOSC-hub services. The pilot will be carried out by the members of the EGI Federation, specifically the EGI AAI technology provider GRNET, in collaboration with Picturae.

JRA1 includes two further pilots for workflow integration, one to test the replacement of an institutional system for loans and visits with ELViS, and one for integration with the DINA collection management system. The aim of these pilots is to develop specifications for developers to integrate other systems and software with ELViS.

ELViS requirements and use cases will be collected across VA and TA, through two partner stakeholder workshops, and by involvement in testing as part of code sprints during the agile development process. Partners will provide their time in-kind for information analysis and testing. Incorporating a large number of partners in the testing process will ensure compatibility with the wide variety of collection management approaches across SYNTHESYS+ partners.

JRA2 - Collections on Demand

SYNTHESYS+ partners have been at the forefront of opening up access to collections both in terms of processes and technological advances. However, **we are still facing a challenge to manage demand and break down remaining barriers to collections access**. JRA2 will address this by developing sustainable mechanisms to prioritise Collections on Demand requests and resolve the technical barriers that prevent certain specialist collections from being made accessible. These mechanisms will be in the form of technological services to support VA to NH collections. The primary focus will be on providing services for 3D digitisation and for DNA sequencing. Cost models, new data pipelines and standard workflows will be expanded and developed for these activities, alongside novel molecular lab protocols to enable large scale DNA sequencing of NH collections. Standards and guidelines will be developed or enhanced in close association with NA3 and NA4. Activities will feed into the VA programme so that in the later VA calls, users can benefit from ongoing developments.

This work package is organised into three components:

1. The **Digitisation as a Service model**, which will support delivery of VA to NH collections. This will provide an effective mechanism of prioritising VA requests. From year two, there will be two calls for VA proposals from researchers, providing time for procedures to be established in year one and promotion of the VA programme. Proposals will be prioritised based on criteria established in year one. Once the process and governance framework are developed they will be integrated into ELViS for subsequent use in managing SYNTHESYS+ VA calls.
2. **Digitisation on Demand (DoD)**. A process will be developed to offer digitisation on demand to users with a focus on 3D tomography models. The services will give online access to institutional 3D collections, working with SMEs such as Sketchfab to build on existing pipelines and web services within national infrastructures including LifeWatchGreece RI and GFBio. 3D methodologies will be based on the

best practice handbook produced within SYNTHESYS3. Data generated from the VA calls and tracked using the ELViS platform will follow the standards defined by NA4 and will be incorporated into the Specimen Data Refinery (JRA3) and used for training (NA2).

3. **Developing the protocol infrastructure for DNA sequencing-on-demand (DNAoD)**. This work builds on a previous review of the current state of the field for sequencing preserved NH collections undertaken during SYNTHESYS3. **The outstanding research challenge is optimisation of protocols and workflows while also making them routine, cost-effective and scalable**. Effective and efficient protocols will be developed to overcome the degradation and low concentrations of DNA in many NH specimens. New developments in sequencing platforms and technologies as well as recent progress in molecular biology protocols, will be tested and deployed to boost accessibility of the genetic and genomic data of preserved collections.

JRA3 - Specimen Data Refinery

The vast majority of Europe's NH specimens either lack any form of digital record, hampering access, awareness and potential use. While some collections are scaling up the rates of collection digitisation, **the community as a whole lacks the digital tools to support these efforts and reduce the cost of doing so**. The Specimen Data Refinery (SDR) will address this need by creating a toolkit based on new artificial intelligence approaches, such as computer vision, data mining, reconciliation services and machine learning, to rapidly enhance minimal NH specimen records. This draws on images of handwritten and typed specimen labels, along with handwritten/typed entries in collection registers. These records are the primary source of metadata about each specimen and at present must be manually transcribed as part of any digitisation effort. JRA3 addresses this major bottleneck in global collections digitisation efforts. The SDR toolkit will be largely automated and built using a microservices model, in which each service can be linked together into workflows that suit particular collections objects. Through the platform these services can be configured through the programmable interface to the SDR (the API), enabling SYNTHESYS+ partners to adapt and apply the services to their particular institutional needs.

In order to build the SDR, JRA3 will:

1. Evaluate and assess the existing landscape for data sources, services and contemporary technological approaches, linking with the work on stable identifiers (NA4.2);
2. Develop a new toolkit composed of optical character recognition, data mining and linkage, image analysis, georeferencing and human interaction services;
3. Build a cloud platform for the toolkit to support the High Performance Computing requirements of the services and enable authenticated access for the consortium; and

4. Ensure the data can be successfully used and integrated with institutional systems and third parties through training courses, workshops and documentation (with NA2).

The proposed software development will build on existing packages and platforms for workflow development, including industry leaders in natural handwriting recognition and optical character recognition, as well as experts in computer vision and workflow development. The work package goal is to automate the semantic classification of elements on in the specimen images (e.g. taxon names, geolocation data, dates), and triage these for appropriate methods of additional processing, where necessary flagging these for manual transcription (so-called human-in-the-loop processing) where automated methods fail to provide sufficient levels of transcription precision.

1.3.3. Access Stream

TA1 - Transnational Access

SYNTHESYS+ aims to provide a minimum of 7,017 TA user days through four annual competitive open calls for proposals, in which only the highest scored applications will be funded. The 21 collection-holding beneficiaries will be grouped into national hubs, forming 13 Taxonomic Access Facilities (TAFs, for example all five German TA beneficiaries will together form DE-TAF). Each TAF will have its own leader who serves on the Project Council to facilitate the effective delivery of the TA. The 13 national TAFs comprise the best collections in each country and national User Selection Panels (USP) of international experts will be held to ensure in-country integration of the offer. Each TAF will aim to include a previous SYNTHESYS TA user on their USP. The national TAFs are based on the largest and most comprehensive collections in Europe, supported by skilled curators and used by highly productive research departments. All TA beneficiaries, both those included in previous SYNTHESYS projects and those newly joined in SYNTHESYS+, have a proven track record of attracting a large number of productive international researchers.

A promotional campaign will begin in M1, featuring all of the 21 collection-holding institutions and drawing potential researchers' (users) attention to the SYNTHESYS+ funding. Users will apply through a single entry point for access to over 490 million specimens, related scientific expertise and associated analytical facilities. The inclusion of new TAFs in SYNTHESYS+ (Israel: IL-TAF and Finland: FI-TAF), and additions to existing TAFs (DE + BE), increases the offer to users compared to previous years. **Potential new users will be encouraged** by including an adapted scoring system with more weight given to applications **focused on societal challenge areas** (for example "Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the Bioeconomy;" and "Climate action, environment, resource efficiency and raw materials"). In the past a number of SYNTHESYS TA applications have received a lower score due to the proposer's lack of proposal writing experience, rather than any deficiency in the quality of their project. The SYNTHESYS+ TA will therefore offer a broadened helpdesk function specifically to support new users and to offer guidance in proposal-

writing. This helpdesk will be clearly separated from the peer review process and the USPs.

Integrating JRA with TA: Improvements to the quality and amount of NH specimen records and metadata in virtual collections as a result of the JRA3 Specimen Data Refinery will greatly improve TA provision. Digital access will enable potential visitors to make better informed choices and plan targeted, shorter visits. TA users will be able to use the improved technologies developed in the JRA2 and JRA3 to digitise collections during their visits – copies of which will be stored by the host institution to benefit future users both within and outside the institution. Successful TA user applications that may contribute to JRA2 Imaging-on-Demand development will be highlighted by the USP to the JRA2 team to allow potential integration, benefitting both the JRA2 and the user via feedback and refinement.

TA users will be asked for their feedback on the development of the JRA outputs, for example on the functionality of the DoD service and the usability of the Specimen Data Refinery tools for research purposes. This feedback will be used to refine the JRA outputs to better meet user demands.

Integrating NA with TA: The harmonisation of policies in NA2 will raise the standards of collections preservation and accessibility, thus ensuring maximum accessibility by both current and future TA users. In addition NA2 will identify areas of unrealised weakness in the management of European collections through the continued and expanded use of collections assessment tools and via user feedback from the TA online system (later ELViS) regarding the quality of access at TAFs. NA2 tackles both TA and VA needs by supporting the community in acquiring digital (data) skills and competencies that enable users to navigate collections information effectively, as well as optimise access and utilisation of NH collections. TA users will be made aware of the protocols developed in NA3 and NA4 via the application process and their clarity and usability will be beta-tested by project beneficiaries and TA users before being promoted to NH collections outside the consortium. NA4 work on digital standards and processes will increase the interoperability of collection data and services, speeding up services delivery and the discovery of data for users. NA5 will engage with the TA user community by seeking TA user representatives particularly in expanding new user groups, including attendance at the planned NA5 workshops.

VA1 – Enhanced Virtual Access

The increased effort by NH collections to digitise their holdings has meant a change in the behaviour of the user community. Once only accessible by physical visits, these collections are increasingly being requested as groups of digital objects, but the low proportion of these collections currently digitised means such requests are currently dealt with in an ad hoc manner by individual curators, rather than in an integrated way across multiple institutions. As noted in section 1.2, the implementation a programme of prioritised VA across the SYNTHESYS+ network will establish common standards and practices, allow

transparency for users, and enable user participants to assist collections in digitising their material.

SYNTHESYS+ will offer an integrated DoD service through VA. Nineteen collection-holding institutions will be freely available to the global user community through two open calls, held in years 2-3. Users can request resources that they wish to use in digital format. Requests will be accompanied by a scientific case and outline of anticipated impact and outputs. An independent committee will be used to prioritise requests for VA through the associated JRA task 7.1 and top ranking proposals will be fulfilled, undertaking the necessary digitisation, data curation and provisioning necessary to make the data freely accessible through open public portals. Each institution receiving a user request for DoD will have a VA Institutional Coordinator to promote the VA activities, help to assess the feasibility of digitisation requests to aid the prioritisation process and coordinate the digitisation activities including open access delivery of the data. This activity will make use of an existing network of public Portals (principally GBIF and institutional data portals) to provide unfettered access to collections data and associated media.

All VA requests will be publicly tracked and monitored through a data dashboard, with metadata (minimally specimens digitised, downloaded, cited and the number of publications) submitted to an external board for assurance and to provide value for money recommendations. This ensures institutions and users make best possible use of the service and provides critical information for a wider rollout of VA envisioned through DiSSCo.

Provision of greater and more flexible VA will exploit the existing digitisation capital equipment within participating institutions, with EU financial support only being used to cover the technological and scientific support activities needed by researchers to effectively use this service and make the data openly accessible.

Integrating JRA with VA: Initial VA provision will rely on digitisation workflows already in operation at participating institutions, but as JRA2 work progresses to support 3D and DNA sequencing-on-demand (DNAoD), we expect to make these new workflows available. ELViS (JRA1) will be used to process VA requests and track the citation/use of digitised collections. This is one reason why there will be no VA call in year 1 of SYNTHESYS+; feedback from the first VA requests will be used to refine the ELViS platform. Tools and services developed under the Specimen Data Refinery (JRA3) will be deployed across the consortium to improve the efficiency and speed of data delivery in the second half of the VA programme.

Integrating NA with VA: Work examining the feasibility and development of the additional workflows for on-demand extraction/sequencing requests will be supported by NA3 (task 3.2). Collections data standards developed in NA4 will also underpin VA, in particular the activities associated with the stable identifier framework rolled out to SYNTHESYS+ partners in task 4.2 will ensure source material accessed in the VA can be easily identified and cited. Training delivers through NA2 will be aligned with the needs of VA to improve

Collections on Demand capacity across participating institutions and deliver a consistent, high quality service.

1.4 The Ambition of SYNTHESYS

Knowledge of European (and global) biodiversity is currently spread across hundreds of databases, numerous publications, approximately 2 billion physical specimens and a network of experts. To deliver an organised knowledge base providing the foundations for accelerated taxonomic, ecological and environmental research, and to enable evidence-based policy decisions, all this knowledge needs to be made accessible and continuously curated in digital form as a digital enrichment of the physical objects. This is also required to be able to undertake “expeditions” within the collections themselves, effectively leveraging information already collected, and supplementing more costly field expeditions to collect new objects and data. **Only by combining these data sources can the true value of these datasets be realised, yet for the moment a majority of NH collections in Europe remain digitally invisible and inaccessible for the majority of research and analysis.**

In the United States, the main coordinating efforts regarding natural science collections are carried out by iDigBio, which was envisioned by a Strategic Plan of the Network Integrated Biocollections Alliance (NIBA; now Biodiversity Collections Network, BCoN) in 2010 and subsequently funded by a National Science Foundation program for Advancing Digitisation of Biodiversity Collections (ADBC). The initiative brought together many of the 1,600 biological non-federal collections across the United States, representing approximately 1 billion specimens. Other important initiatives are led by pioneer countries such as Australia and Brazil. In the former, the Atlas of Living Australia (ALA) is an example of a collaborative network that collects aggregated biodiversity information, amounting to over 50 million recorded data items from 201 collections. In the latter, the Reference Centre on Environmental Information (Centro de Referência em Informação Ambiental - CRIA) is a Brazilian focal point for the dissemination of biological information in the interest of the Brazilian science and research community. **SYNTHESYS+ will play a key role in coordinating such efforts in Europe.**

The twenty-year ambition: One World Collection and Modelling the Biosphere

A pan-European RI is needed to solve the current limitations on the access to and use of collections data. This is fully envisioned within the DiSSCo initiative, but **SYNTHESYS+ will fulfil a critical transitional role in preparing the community in support of this next stage of integration**:

One World Collection

A long-standing challenge in life sciences is to document all species on the planet and to accurately place them on the tree of life so that they can be understood and studied in context to all other biodiversity on the planet (David and Taquet 2017). This ‘taxonomic impediment’ currently frustrates both fundamental research on the origins and diversity of

life, as well as more applied efforts to monitor and conserve biodiversity, combat climate change and protect our health and food supply. The data necessary to address this impediment lies in the estimated 7,000 NH collections worldwide. These institutions are characterised by the presence of unique collections that if brought together could provide the best representation of the natural world, its past, present and our prediction of its future. **The One World Collection vision is to bring together the data, routes of access, service provision and aspectsof their operation into a single distributed infrastructure.**

Modelling the Biosphere

Environmental change poses one of the greatest risks to the natural systems upon which our society and economy depend. As human activity pushes the Earth beyond its safe planetary boundaries, we need robust predictive models to make the best decisions for our future – this is the Anthropocene challenge in a nutshell. Current models of environmental change are typically based on relatively crude assumptions about the relationship between geological, biological and atmospheric systems and on data-poor models. Ambitions to integrate global collections and use this as a model of ecosystem function are long term community goals (Purves et al. 2013). These cannot be achieved without further international cooperation, improvements to our infrastructure (especially access to digital collections), advances to our service catalogue and innovation to our systems and processes. **NH collections contain key information on geological, spatial, environmental and ecological changes, throughout recent and deep time, which can be used to build better predictive models for decision-making and management.**

The role of SYNTHESYS+ in fostering a culture of International cooperation

SYNTHESYS+ will close a gap in the global collections infrastructures landscape and thus contribute to a global effort to provide bio- and geodiversity data. Work package leadership by the major global standards and data sharing organisations (GBIF, GGBN and TDWG), coupled with the networking association for the European natural science collections (CETAF) will help sustain a culture of cooperation between project participants and the international community. Working closely with the leaders of comparable international efforts (e.g. iDigBio, CONABIO, CRIA, SANBI and ALA), and building on efforts to implement the high-level vision of the GBIO, SYNTHESYS+ will contribute to the development of community roadmaps for collections, standards, and data infrastructures necessary to integrate our collections (moving towards the vision of One World Collection) and use these data as a model of ecosystem functioning (Modelling the Biosphere). Contributing to these efforts will be new user communities and industrial partners developed through SYNTHESYS+ activities (particularly NA5 task 3), as well as representation within priority areas for geographic collaboration (NA5 task 2). These sectors are typically underrepresented within international community efforts to integrate natural science collections and data. In this respect SYNTHESYS+ will take a leading role in establishing and implementing a coordination mechanism for stakeholders internationally to identify, agree and plan shared priorities for an interconnected collections RI. Implementing such a mechanism will allow SYNTHESYS+ and other regional and national

initiatives to benefit more effectively from distributed international investments. As a truly international undertaking, this is essential to achieving long term community ambitions.

Access to an integrated state-of-the-art collections infrastructure

SYNTHESYS+ will unify physical and digital access to European collections, and will work toward providing linked collections data, driving policy and process harmonisation. Through new forms of collection access, aggregation and linkage of European collections, critical new insights will enable scientists to address some of the world's greatest challenges. Our community ambition is to provide the data and tools to support consistent and comprehensive global discovery and use of information from all sources about the biodiversity of any defined area over time, covering all taxonomic groups. Within SYNTHESYS+ the development of ELViS, providing a unified metadata catalogue of European collections, combined with efforts to increase standards compliance (e.g. Linked Open Data capabilities, CETAF stable identifiers and IIIF metadata standards) across the institutional data repositories, is a major step toward achieving this ambition. This will create a mechanism that enables users to locate collections and data in a language that they can understand, and aid the construction of comprehensive catalogues that reuse authoritative collections data. In the long-term, this will allow the community to work toward all species information being managed and curated as an inter-connected digital knowledge base.

In addition to our collections and data infrastructures, **SYNTHESYS+ has the potential to have a transformative impact on collections infrastructure facilities**. SYNTHESYS+ collection partners hold a wide range of complementary instrumentation which is opened up to shared access through our TA and VA programme. Methodologies like computer tomography, scanning electron microscopy, ultramicrotomy, X-ray fluorescence and X-ray diffraction, various kinds of light and fluorescence microscopy, 3D imaging, 3D printing, mass spectrometry, organic chemistry techniques and automated molecular biology labs are made available to SYNTHESYS+ user communities, enabling them to conduct excellent research of the highest quality. Ultimately, DiSSCo envisages greater consolidation of this instrumentation, improving quality through centres of excellence concentrating on specialised techniques. Some degree of duplication will always be necessary due to the limitations associated with moving fragile collections. Nevertheless, the construction of an integrated instrumentation catalogue within ELViS will help to minimise the duplication of instrumentation across the consortium

Provision of cutting edge services

SYNTHESYS+ will undertake Joint Research Activities necessary to deliver three distinct classes of service that will transition to ultimate management by DiSSCo, aiming to bring together mobilisation (e.g. SDR), access (e.g. ELViS) and support (e.g. helpdesk, dashboard and training) services for the NH collections community.

Fundamental to achieving the ambition of DiSSCo is the means to deliver vast quantities of digital information from European collections. For every year of the DiSSCo programme,

the community aim is to move, digitise and transcribe data from 40 million specimens through industrial processes. Joint research undertaken through SYNTHESYS+ JRA3 is intended to develop the technical processes necessary to support this level of data mobilisation. Using advanced machine learning processes employing computer vision and artificial intelligence methods, the Specimen Data Refinery (SDR) is intended to enhance minimal NH specimen records using images of specimens, their labels and collection registers. This platform will also provide human-in-the-loop functionality thus allowing experts and members of the public to improve the automated processes and enhance records (e.g. via crowdsourcing). This cutting-edge innovation, the basis of which was established through work in SYNTHESYS3, will have a transformative effect on the collections community, putting within our grasp the community ambition of having a complete catalogue of European NH collections, used and enhanced by multiple user communities.

Fulfilling the DiSSCo vision for VA requires SYNTHESYS to go beyond the construction of digital collections catalogues to develop protocols for high-throughput 3D imaging and DNA sequencing, increasing access to the phenome and genome of NH specimens. The development of processes and pipelines to service requests for 3D models of NH specimens (DoD) and the protocol infrastructure for DNA sequencing-on-demand (DNAoD), is the subject of JRA2. This work package aims to raise the current Technology Readiness Level from TRL3 (experimental proof of concept) to at least TRL7 (system prototype demonstration in operational environment), such that high throughput protocols are in place to support a full range of digitisation services. Having developed the necessary content and technical infrastructure to deliver a “Collections on Demand” service, SYNTHESYS+ will need a mechanism to triage requests for digital access to collections. This is the subject of JRA2 task 7.1, which will develop and operate the criteria necessary to prioritise VA requests, as well the governance framework necessary to manage this process within ELViS. This supports the DiSSCo ambition of centralising access requests and will be applied to service the VA programme of SYNTHESYS+.

Innovation potential

The efforts of SYNTHESYS+ in the development of mass digitisation processes, especially the computer vision and machine learning activities through the Specimen Data Refinery, will lead to technical innovation in areas of digitisation, data/media processing and artificial intelligence techniques for text, handwriting and image recognition. A patent search was carried out as part of the DiSSCo preparatory work in 2017 and no relevant results for mass digitisation were found. **Working with our SME partners, SYNTHESYS+ innovations will promote industry solutions in these areas, responding to new requirements with new products and services.** Underpinning this will be advances in technical interoperability standards and collection data exchange standards.

Unifying access to the digital and physical collections will ultimately generate an immense knowledge base that puts Europe at the forefront of supplying authoritative expert information about the natural world. The knowledge graph services planned within DiSSCo and initiated within SYNTHESYS+ through standardisation activities, will accelerate

development of e-infrastructure services in Europe. This work will promote federation of European environmental RIs by providing the reference data required to operate their services and stimulate innovation drawing on the use of big data technologies. This work will also have the effect of creating a critical mass of highly skilled researchers focused on major scientific and related societal questions across three main areas:

1) Biodiversity and natural resources underpin contributions to the production of food, cloth, building materials and medicines, as well as the provision of sustainable energy, rare minerals, and ecosystem services. An integrated European collection RI will provide new ways to study the chemical and physical properties of biological and geological materials and resources. **Related to this, cooperation organisations such as the European life-sciences Infrastructure for biological Information (ELIXIR) will provide scientists new and sustainable ways to develop and design such materials.**

2) Economic sectors like land-use planning, environmental assessment and health are strongly connected to biodiversity and natural resources. One of the fastest growing market sectors covers science-based activities for sustainable management of our natural environment, including energy production, mining, agriculture/forestry and other areas of bio-economy. In addition to these private sector industries, public authorities spend considerable resources on regulating and managing biodiversity and the wider natural environment. Advances underpinned by SYNTHESYS+ have significant potential to benefit these sectors by providing new scientific knowledge on the properties and dynamics of our bio- and geosphere.

3) Biodiversity and ecosystems are crucial for buffering (extreme) environmental changes. There is a fast-growing interest in innovative infrastructure services for early warning indicators and natural disaster mitigation. SYNTHESYS+ has the potential to aid scenarios for decision-support and environmental management.

2. Impact

2.1 Expected impacts

2.1.1. Lowering access barriers across borders and disciplines

Building on the transformative work of previous SYNTHESYS projects, SYNTHESYS+ will act to broaden the direct beneficiaries of natural science collection information, both geographically and across disciplines, by introducing and fully supporting VA alongside the (physical) TA pillar of the project. TA typically requires deep and thorough understanding of specimen handling protocols and is usually performed by expert personnel. Though this is a key part of biodiversity discovery, it is of lesser value to non-biodiversity experts and specifically to non-taxonomists. Mobilisation (digitisation, annotation and publishing) of information extracted from collections can significantly increase the number of beneficiaries. The demand for VA to collections information has increased significantly over

the last few years. Whilst TA across European natural science collections accounts for 16,000 researcher visits per year (approx. 150,000 access days), visits to existing data portals providing VA to European collections are at least a magnitude higher.

SYNTHESYS+ focuses on promoting a trans-disciplinary approach to data use. By fully adopting FAIR principles and translating them into actionable policies across the network of partners, SYNTHESYS+ puts natural science collections information at the heart of data-driven scientific practices.

The SYNTHESYS+ access programme will balance access modes (physical/virtual) based on the scientific needs of wider communities of practice. As such, it will prioritise content generation based on urgent scientific needs and provide open and free VA to relevant data.

2.1.2. Directly contributing to the European Research Area

Digitisation requires new skills and capacities from old professions and institutions, to which the EU has responded through the conclusions on cultural heritage as strategic resource (EU Council 2014). This report gives an EU perspective on why natural science collections, including digitised collections, are cultural heritage and have strategic and economic importance to Europe as a whole. The progress of digitisation provides opportunities for new services which create employment, are inclusive, and have global reach. Digitised data can be integrated with other data for a trans-disciplinary approach, allowing for the linking of expertise from different domains, and fostering the creation of new capabilities, as required by EU economic policies.

During the SYNTHESYS3 project, a series of use cases were identified (Kvaček et al. 2016) that highlight the feasibility of combining data from natural science collections with data across disciplines and fields of science in order to advance scientific knowledge in areas relevant to urgent challenges. These show collections data being used in fields as disparate as human health, invasive species, and urban greening, with impact on a similarly wide range of end-users.

SYNTHESYS+ will directly contribute by delivering components of a much needed pan-European RI, refining and harmonising data management practices across the participating organisations and promoting open access to biodiversity information. By investing in multi-modal access (physical and virtual) and by developing added-value scientific services that allow researchers to further benefit from the collections data availability, SYNTHESYS+ will contribute to the development of an urgently needed online toolbox - a set of services that significantly lower the entry barrier for new and diverse users to create, access and interpret collections information.

2.1.3. Novel virtual access services and support structures

SYNTHESYS+ builds on the trust relationships already built between service providers and users to develop training modules and unified support mechanisms. Such mechanisms will be handed over to DiSSCo RI to ensure their persistence. The project employs innovative ways of creating new, open access scientific content through digitisation activities that are focused and synchronised across the TAFs. By putting scientific needs at the centre of how the consortium generates new information for VA, through open calls for submission of digitisation requests (DoD) that are independently prioritised on their scientific merit and urgency, we will ensure the most impactful collections are identified for digitisation.

To fully optimise access to European collections, it is imperative that TA and VA are not disconnected modes, but rather complementary, providing more holistic ways of retrieving information from European scientific assets. To make this possible, **SYNTHESYS+ will develop a novel service, providing a unified (one-stop shop) entry point for access requests and visit monitoring**. The introduction of ELViS will significantly lower overheads for both users and administrators. Post-project operation of ELViS is ensured by the DiSSCo RI consortium. Working with the consortium on highly specialised aspects of work will allow **SME partners to innovate in new areas with a variety of institutions, improving their competitiveness in these and other areas post-project**.

2.1.4. Policies, processes and standards harmonisation across organisations

Technical innovations like high-throughput Next-Generation Genomic Sequencing (NGS), and large-scale digitisation facilities, including 3D imaging, rapidly increase the volume of research data. To fully benefit from this, existing strategies and protocols for sustainable data annotation, storage, availability and analysis must be improved. Recent changes in legislation (e.g. Regulation [EU] No 511/2014, concerning the implementation of the Nagoya Protocol, or [EG] No 116/2009, regarding the export of cultural goods) increase the need for traceability of genetic resources and for meeting documentation requirements. Thus, practical tools for collection-related activities are needed to integrate those tasks into the day-to-day work of researchers and to maximise usability of the data generated from both current and future collections. Common standards are urgently needed to develop shared documentation policies. For example, to respond to Regulation (EU) No 511/2014, collections must elaborate common best practices for documenting and utilising newly acquired biological material. New legal instruments require institutions to develop appropriate internal mechanisms and procedures that:

- record the terms and conditions under which biological material are accessed or otherwise acquired;
- record relevant information on access and utilisation of biological material and the benefits arising from that utilisation;
- record transfer of biological material to third parties and develop common standards and syntax for transfer of that material between collections;

- record when and how biological material pass permanently out of custodianship (consumption /disposal).

SYNTHESYS+ is introducing two complementary work packages to address the gaps in existing standards around managing molecular (NA3) and digital (NA4) information. **These specifically address the legislative and policy aspects of managing collections within these domains and are led by international organisations in their respective fields. Both will develop a more complete corpus of standards to be incorporated into organisational policies and practices. This development and use of these common standards and protocols will maximise efficiency, raise transparency and increase interoperability.**

2.1.5. Positioning European organisations in a global context

The volume and international significance of European NH collections, positions Europe at the very heart of discovery and understanding of the natural world. Previous SYNTHESYS projects have provided considerable support in unlocking the scientific value of these collections. Challenges around TA and VA to collections are shared across continents. In the USA, comparable initiatives such as ADBC (and the iDigBio hub) are trying to address issues around mass-scale collections digitisation and access. Similar initiatives are now in place in China, Australia, South Africa and elsewhere. Building on past progress, European organisations are now well-positioned to leverage their unique collections and position themselves as the driving force in developing a global research infrastructure, which will sustain a set of common resources and tools at a global scale.

A dedicated work package (NA5) will be focused on internationalisation and the expansion of our user community. This will serve to strengthen the European cohesion of relevant stakeholders, allowing them to better position themselves in the global landscape

2.1.6. Developing robust training programmes

Lack of capacity to engage with the practices and benefits of open science is considered one of the most important bottlenecks reducing the uptake of e-science. As SYNTHESYS+ invests in a balanced access model, combining TA and VA access, it will also work towards capacity enhancement activities around digital skills and competencies. By leveraging the experience of partners such as CETAF and the affiliated Distributed European School of Taxonomy (DEST), SYNTHESYS+ will further refine existing training programmes organised across Europe, strengthening digital training. As these programmes already award academic credits to participants, when applicable, we will provide better incentives to graduate students to increase attendance. At a professional continuous education level, SYNTHESYS+ will harmonise training/vocational programmes across participating organisations. This will ensure that a minimum level of training is provided to professionals across the partnership. **After the completion of SYNTHESYS+, these harmonised modules will be adopted by DiSSCo, thus ensuring their continuous update and further development by DiSSCo facilities.**

2.1.7. Fostering innovation and science-driven data interoperability

Innovation within SYNTHESYS+ will be fostered through a reinforced partnership of research organisations with industry. JRA1-3, under the coordination of the dedicated JRA stream lead, will focus on developing novel solutions for lowering access barriers to collections and streamlining physical and VA. Across these activities SYNTHESYS+ will activate partnerships with industrial partners in the areas of computer-assisted tomography, 3D imaging, industrial scale digitisation, big data management and cloud-based end-user service deployment.

Mobilisation of collections information presents great scientific opportunities but not without significant additional cost. Thus prioritisation of content generation, annotation and publishing will be at heart of how SYNTHESYS+ approaches VA across Europe. The United Nations Sustainable Development Goals and their corresponding Targets underpin the overarching scientific drivers that set SYNTHESYS+ priorities for new content generation. The VA programme of SYNTHESYS+ will specifically act under the umbrella of those urgent challenges, informing how the consortium prioritises virtual content generation, annotation and publishing. Furthermore, the open calls process will allow the consortium to better understand the practical needs of the community, and tosupport better cross-disciplinary linking of datasets.

2.1.8. Leading to a Distributed System of Scientific Collections: a pan-European RI.

As an advanced community, the SYNTHESYS consortium has been working in the SYNTHESYS3 project (2013-2017), towards designing a pan-European RI that will streamline access to collections and transform a fragmented model into an integrated European system. The Distributed System of Scientific Collections (DiSSCo - http://dissco.eu) proposal for a new pan-European RI was accepted by the ESFRI (http://esfri.eu) for inclusion in the 2018 roadmap update of European RIs.

As set out at section 1.3(a) above, the DiSSCo preparation and construction programmes will lead to a fully operational new RI by 2025. SYNTHESYS1-3 have been pivotal in building a socio-cultural, governance and technical consensus around DiSSCo. In this respect SYNTHESYS+ will act as a critical transition project, undertaking the groundwork to migrate a loosely joined network of access providers into an integrated and self-sustainable system of European collections.

Specifically SYNTHESYS+ will be instrumental in: (a) networking activities, strengthening European and international stakeholder engagement; (b) innovation activities, delivering a new single entry point system for requesting and managing TA and VA (ELViS); (c) standards development activities, facilitating intra- and crossdisciplinary data interoperability; and (d) support and training activities, improving future user audience engagement with DiSSCo services.

2.2 Measures to maximise impact

a) Dissemination and exploitation of results

Plan for dissemination and exploitation of the resultsThe long-term success of the project and the impact of the work achieved rely on the development andimplementation of a comprehensive communication strategy. The existing strength of the network of NH collectioninstitutes, both within Europe and internationally, will be key for ensuring that the activities of SYNTHESYS+ willbe successfully communicated and taken up beyond our immediate network. In order to facilitate this, the projectincludes partners that have an existing role in bringing the whole of our community together. GBIF, aninternational open-data RI, will be leading NA5 which will focus on the dissemination and internationalisation ofthe project. CETAF is a research and collections network with 33 members representing 59 institutes from 21countries. CETAF's leadership of NA2 provides a critical platform to share the results and services developedwithin SYNTHESYS+. TDWG is an international collaboration to develop, adopt and promote standards andguidelines for the recording and exchange of data about organisms. As a partner in the project leading NA4, thisassociation provides an international communication channel with NH collections worldwide. The inclusion of GGBN as an international network of institutes providing a platform for biodiversity biobanks to preserve genomicsamples will provide NA3 with the means to disseminate results and services to 59 biobanks from 24 countries.

In widening the dissemination further outside our community we need to tailor strategies to different clearly identified target audiences. The relevance of NH collections to society as a whole has been clearly identified, but there is still a need to work with focussed sections of society to ensure that a wider engagement and understanding is achieved. NA5 activities will extend the potential to work with the **Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)**, to reach out to the policy-makers in 128 member states. This will also provide a forum for communication with a large number of NGOs, organisations, conventions and civil society groupings. Additional groups of end users, including industry, commercial enterprises, local authorities and education facilities will also be identified and dissemination strategies designed to reach these stakeholders.

In building the dissemination strategy, this project will develop the work carried out within SYNTHESYS3 in creating a **European Roadmap for NH collections** (Kvaček et al. 2016. This roadmap provided a review of 26 use cases, backed up by documentary evidence to demonstrate how collections, expertise and services can be used to find solutions to a range of societal challenges addressed in the European Union H2020 Framework Programme.

Within SYNTHESYS3 a study on the emerging new uses of NH collections prioritised the need for 1) the identification of potential new user groups for NH collections; 2) the need of maintaining contact with priority user groups and improve dissemination on the value of collections within these communities; 3) the co-development of collection-based services with new and priority user groups (specifically DoD, sequencing, environmental sample

vouchering, isotope reference samples, collection management services, ID services, 3D print on demand and collections hosting); and 4) the development of a more customer-oriented approach to service provision. Actions in SYNTHESYS+ are a direct response to this study.

How the proposed measures will help to achieve the expected impact

The inclusion of GBIF, CETAF, TDWG and GGBN, as permanent organisations with strong membership from within the SYNTHESYS+ partners, will ensure that the results and services developed within the project continue to be taken up by existing members and reach new members beyond the consortium. The role of NA5 with an explicit focus on internationalisation will magnify the impact of our activities beyond Europe, so the actions of SYNTHESYS+ becomes a global endeavour, reaching new user communities and with a focus on EC priority areas. Sustainability is assured through the DiSSCo framework, and the actions of JRA1 (ELViS) will ensure that access opportunities are broadly advertised and tracked. In particular, VA has the potential to transcend the practical limitations of traditional physical access to collections, reach new user audiences who can work across multiple collections as part of a single access request, and thereby feed directly into new use cases for collections.

Management of research data generated during the project

Open science agreements will be put in place at the start of the project with all partners. Gold standard open access to articles, data and other published documentation will be the standard dissemination procedure to maximise outreach and impact. All the software developed within SYNTHESYS+ will be made publicly available, with the source code published under an appropriate free, open source license. As with earlier iterations of SYNTHESYS, all outputs will be “open by default”, with exceptions only granted where there are agreed ethical, privacy or pre-agreed commercial interests. A **Data Management Plan (DMP)** for SYNTHESYS+ will be available at M3 as a public deliverable. In addition, the DiSSCo draft DMP will be available at M19. These plans will describe an architecture that will ensure that these data fulfil FAIR requirements, use internationally recognised standards, and be deposited in an accessible repository.

SYNTHESYS+ partners as well as users of the Access program will have to respect rules and procedures to deposit data in the RI. They will have to follow the established standards to describe data and use standard, interoperable file formats and packaging standards. In collaboration with the relevant standard bodies (e.g. TDWG, GGBN), the standards themselves will be scrutinised to establish how far they fulfil the needs of the new RI, and where updated are needed. Moreover, the methods currently used in the digitisation processes will be analysed in NA4 and recommendations and suggestions on workflows and software deliverables will be outlined to facilitate quality assurance of digitisation. Additionally, JRA3 will contribute with the selected methods for automatically analysing and annotating data from images, so that digitised data can be made discoverable and the need for human effort in converting data from images into structured machine-readable form can be minimised.

Openness and sharing have long been central to the operation of NH collections. This community has been implementing open access to electronic data since 2001 when global initiatives such as GBIF became operational. However, approaches to dealing with exceptions to the “open by default” principle, need further analysis. For example, the community needs agreed methods to ensure that collection localities of endangered species are not disclosed. In this respect international legislation often differs outside of Europe, creating discrepancies in the way data is published. These legal and policy aspects will be considered under NA2 and NA5, providing an international picture that will contribute to framing the global communities common research agenda, and help identify a pathway to align individual strategies.

Knowledge management and protection

A collaborative, virtual workspace (Teamwork, under the auspices of http://dissco.org) will be used to exchange information, results and deliverables. This will be established for project, document and product (e.g. software) management activities at the beginning of the SYNTHESYS+. Guidelines for the use of Teamwork will be developed to include quality control procedures as well as the operation for external users outside the SYNTHESYS+ consortium, who may be involved in contributing select materials. Many SYNTHESYS+ deliverables will generate technical reports that will remain available within the document repository. In some cases these will be “live” documents that will continue to be edited after the formal SYNTHESYS+ task is complete. The Teamwork workspace builds on the SYNTHESYS+ preparatory activities conducted within the system by the consortium in developing this proposal. This cloud-hosted workspace will be transferred to NHM London at the end of the SYNTHESYS+ project to guarantee its availability and long term archival.

The maturity of the collaboration within the SYNTHESYS+, coupled with our long-standing partnership with networks such as CETAF and GGBN, demonstrate the ability of this community to manage the sharing of knowledge whilst protecting the development of innovations. The consortium will exploit these networks to the fullest extent possible to maximise the dissemination of these outputs. In addition, this community has the necessary legal experts within the consortium to deal with any innovations that may need protection. Where necessary, this expertise will be brought in to develop the SYNTHESYS + Consortium Agreement.

b) Communication activities

To expand the network and engage general public and industry, efficient communication is essential to share developments made during SYNTHESYS+. In particular, the project needs to highlight the innovations and societal relevance of SYNTHESYS+ activities, expressed as part of the wider ambitions of the DiSSCo initiative. While traditional dissemination activities, such as flyers and conference presentations continue to play a role in dissemination, these alone are insufficient to significantly broaden our user base. Consequently, SYNTHESYS+ and its partner networks (CETAF, GGBN, TDWG and GBIF) will undertake extensive networking actions using a range of innovative approaches to communicate our activities and reach beyond our traditional user community. This involves

leveraging the brand of the larger consortium partners to undertake high profile communication on behalf of SYNTHESYS+. A small number of these approaches, such as participation in TV documentaries and the production of high quality promotions videos, were tested during SYNTHESYS3 and generated high levels of engagement with new audiences. We expect to develop these further to expand the reach of SYNTHESYS+.

Potential activities include:

- Participation in news and other broadcast media (e.g. TV documentaries)
- Web multimedia videos
- Press and magazine articles
- Blogs and improved use of social media
- Engagement of high profile external speakers within SYNTHESYS+ events
- Conference presentations to new user communities
- Policy focused events with national government and EC representatives
- Pop-up events at exhibitions and technology fora
- Engagement in public science festivals

As part of this work, **SYNTHESYS+ will make greater use of our user community to showcase the breadth and relevance of their activities**. In particular we will focus on VA users to highlight the new research opportunities associated with digital collections. To further strengthen communication the SYNTHESYS+ stream coordinators will help to identify highlight activities from within their respective streams and disseminate best-practice in outreach. They will also play a vital role in strengthening internal communication, especially across work packages and with external initiatives beyond the consortium.

Several work packages have a special role in supporting communication. NA2 through the CETAF network will ensure effective delivery and dissemination of best practices and standard operating procedures internally across the consortium, providing helpdesk support and developing support resources. However, NA5 takes on particular significance in communication activities through a programme of workshops. These will convene a range of SYNTHESYS+ partners relevant to workshop topics and will fund the engagement of a selected set of attendees jointly to develop the resulting roadmap recommendations document. Participants will include European stakeholders, as well as international initiatives including Catalogue of Life, Biodiversity Heritage Library, International Barcode of Life and other relevant international institutions and programmes. Examples of the latter include iDigBio from the United States, ALA from Australia, SANBI from South Africa and CONABIO from Mexico. This approach will improve the quality and significance of the developed roadmaps and will enable improvements to the efficiency of the European biodiversity infrastructure landscape. Workshops will also consider issues such as language barriers to support smaller NH collections working in national languages. SYNTHESYS+ will take a leading role in establishing and implementing a coordination mechanism for stakeholders to identify, agree and plan shared priorities for an interconnected collections-focused RI. **Implementing such a mechanism will allow**

SYNTHESYS+ and other regional and national initiatives to benefit more effectively from distributed international investments.

3. Implementation

3.1 Work plan – Work packages, deliverables

SYNTHESYS+ will consist of 11 work packages (WP) divided into three streams, representing the three key activities: Networking Activities (NA), Joint Research Activities (JRA) and Access (TA and VA). WP1 will cover management including promotion, dissemination and communication activities. Four other Networking work packages (WP2-5) will each be led by a global networking organisation covering the work on standards, processes, best practices and internationalisation of the community. Three JRA work packages (WP6-8) will focus on the technical innovations to improve the effectiveness and convenience for users to access NH collections to conduct research. WP9 will cover the TA, providing users with an offer of over 490 million specimens at 21 different institutions in 13 different countries. WP10 will cover the VA, a digitisation on demand service providing virtual collections from 19 collection-holding institutions freely available to the global user community. WP11 addresses the ethical requirements of the project, linking to task 1.4 on Managing responsible and ethical research in NA1.

Networking Activities (NA): Work packages 1-5

NA1 (WP1) will manage the SYNTHESYS+ consortium, ensuring the effective implementation of the work packages on behalf of the EC, the project beneficiaries and the users. Consortium-wide Tasks on promotion, dissemination and communication will be included in NA1, with the central project management team ensuring these tasks are coherently and effectively carried out both within the consortium and externally. NA2 (WP2) will ensure effective harmonisation, delivery and dissemination of best practices and standard operating procedures across the consortium and to the wider NH collections community, led by CETAF. NA3 (WP3) will develop, implement and disseminate standardised best practices to support sequencing and biobanking activities, led by GGBN. NA4 (WP4) will provide the standards to allow technical coordination between institutions that will drive innovation by linking digital data together interoperably, led by TDWG Europe. NA5 (WP5) will ensure integration of major international stakeholders to develop the global collections research agenda, led by GBIF.

Joint Research Activities (JRA): Work packages 6-8

Three JRA work packages will generate technical innovations to serve the user community. JRA1 (WP6) will create ELViS: a platform building on the existing SYNTHESYS online TA application system, supporting Access requests, tracking outputs and ultimately integrating with Collections Management Systems to support loans. JRA2 (WP7) will develop innovative solutions for providing “Collections on Demand” to meet user requests by breaking down current barriers to collections access. The work package will address key

gaps in our technical infrastructure and institutional capacity to undertake and process Collections on Demand requests. JRA3 (WP8) will build a “Specimen Data Refinery”: a platform to integrate artificial intelligence and human-in-the-loop approaches toextract, enhance and annotate data from digital images and records.

Access: Work packages 9-10

Access to collections, their associated expertise and specialised equipment is vital in the field of NH research. The demand for access is demonstrable: 8,626 eligible TA applications were submitted to SYNTHESYS since project initiation in 2004. Users were pan-European, with applications received from 38 eligible European States.

TA (WP9) will be covered by a single work package, using an integrated online application and evaluation system (transitioning to the ELViS platform from year 2 in the project). This is a large work package: the 21 collection-holding beneficiaries are organised into 13 national Taxonomic Access Facilities (TAFs), each with its own TAF leader to facilitate the effective delivery of the TA. SYNTHESYS+ aims to provide a minimum of 7,017 TA user days through four annual open calls for proposals.

VA (WP10) will also be covered by a single work package, offering user-driven Collections on Demand services to specimens at 19 collection-holding beneficiaries. The VA will enable Users to identify priority specimens and/or collections for digitisation with data deposited and annotated in internationally recognised repositories operating FAIR practices. DoD requests will be submitted by users to host institutions to agree feasibility of digitisation. Requests will be prioritised by an external panel facilitated by JRA2. Data will be delivered through openly accessible public data portals, and data will be tracked to demonstrate impact (registered in ELViS, JRA1).

WPs 1-9 will all start at the beginning of the project, with the VA programme starting in the second year following the refinement of the prioritisation process of digitisation requests in the JRA WP7. Project timings are shown in a Gantt chart. There are some further dependencies between the work packages and tasks, as noted in Fig. 6.

Ethics: Work package 11

A specific work package addressing ethics requirements will link with task 1.4 of the management work package (NA1). Activities will be undertaken in a non-EU country (transnational and virtual access services provided by beneficiary 21, HUJI. Research will be related to the Access activities and therefore of the same nature as that undertaken in the EU-based TAFs; therefore the research conducted outside the EU is legal in the EU Member States.

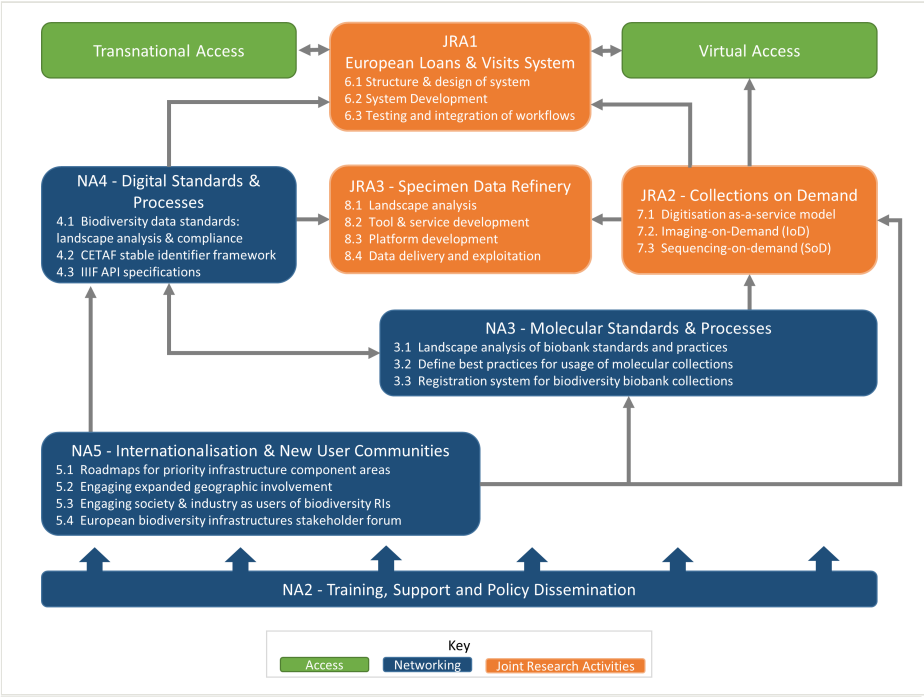

Figure 6. doi

Diagram showing the relationship between the workpackages

Interrelationships of SYNTHESYS+ components

The relationship of the work packages, excluding management (NA1) which provides oversight and governance for the whole project, is illustrated in Fig. 6. Work packages are shown in their three streams (networking, research and Access).

3.2 Management structure, milestones and procedures

Management Structure

The SYNTHESYS+ management structure is illustrated graphically in Fig. 7

NHM Management Team

The project coordinator (PC) is the Natural History Museum London (NHM). In this role, NHM will be in charge of administration and management of the project and acts as the intermediary between the beneficiaries and the European Commission. The project coordinator shall, in addition to its responsibilities as a beneficiary, perform the tasks assigned to it as described in the Grant Agreement and the Consortium Agreement. The coordinator Dr Vince Smith has the overall and high-level responsibility of the project

implementation. The coordinator has the support of a project manager Dr Kristina Gorman who will be the first point of contact for all work package matters.

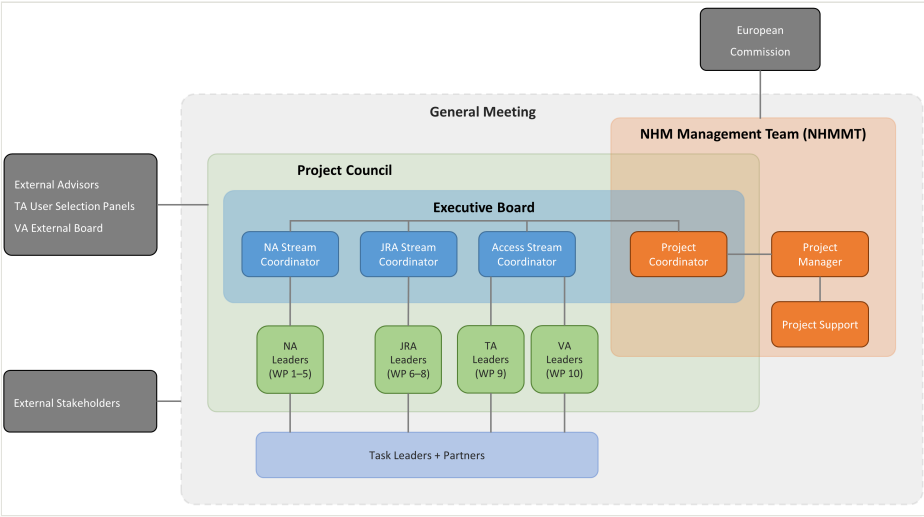

Figure 7. doi

Management structure of the SYNTHESYS+ project

The project manager and a dedicated project support team will assist with the organisation of conferences, workshops and production/dissemination of publications resulting from the work of the project. The project manager will be responsible for managing the project finances by coordinating the establishment of the financial reporting and audit procedures across the SYNTHESYS+ consortium, in line with EU requirements.

The NHM Management Team (NHMMT) will comprise the coordinator, project manager and project support team and will meet at least monthly to ensure that all the components are integrated and communicating effectively and to monitor risks to the Project. NHMMT will maintain an up-to-date risk register, which will serve to alert the consortium to any slippage on deliverables and any under-performing beneficiaries. The risk register will be circulated to beneficiaries every six months.

General Meeting

All beneficiaries (x32) will have a senior representative at the General Meeting (GM) chaired by the coordinator. Annual GMs will be a crucial communication tool and opportunity for face-to-face meetings for the whole consortium. The GMs will be used to ensure the whole consortium is aware of project activities, via presentations on project updates by work package leaders. The GMs will therefore also be important meeting points to review and update the dissemination plan using the full reach of the consortium's network. Any decisions relating to Grant Agreement amendments will involve all beneficiaries.

Project Council

Coordinator (x1), stream coordinators (x3), work package leaders (x10) and Taxonomic Access Facility (TAF) leaders (x13) will meet via teleconference every six months as part of the Project Council. Each work package leader and TAF leader will have a deputy, and deputies will attend as required. These meetings will ensure integration of the three separate streams (Access, NA, JRA) to ensure all potential linkages are secured and exploited within the consortium. It will also be the mechanism by which decisions will be made on Access matters, together with JRA and NA work package leader input to ensure integration between the three streams. The ProjectCouncil will be the ultimate decision-making body of the consortium. Project Council decisions require two-thirds quorum of partners to be present, each council member has one vote, and decisions are taken by a majority of two-thirds of the votes cast.

Executive Board

Coordinator (x1) and stream coordinators (x3) will meet via teleconference every quarter. This group will oversee activities across all work packages and most decisions affecting the day-to-day running of the project will be made by this group. Stream coordinators will ensure integration within the JRA, NA and Access streams respectively, by leading and coordinating the work package leaders within the JRA/NA/Access work packages. Budget and time tolerances will be set at each management level. Task leaders will alert work package leaders if there is a risk their work will deviate from budget, scope or timing. If this exceeds the agreed tolerances, work package leaders will inform the project manager who will advise and in turn escalate to stream coordinators if required. Streamcoordinators will report any risks, issues or deviations at the quarterly Executive Board meetings, and matters will be escalated to the Project Council if required. Milestones are set for each work package to ensure task activities and progress on deliverables are continually monitored and remain on track.

Together the NHMMT, Executive Board, Project Council and General Meeting form a management structure suited to the scale and complexity of the project. The project work will be coordinated via streams to ensure no work package is working in isolation and that common practices are applied across the work packages within the streams (for example, common practices by all TAFs when promoting, administering and approving Access requests). Matters relating to only one particular stream will still be discussed by the main decision-making body, the Project Council, to ensure all streams are integrated.

Management procedures

The NA stream coordinator will be a member of the DiSSCo coordination team based at Naturalis, utilising their widespread network of contacts built up over many years and particularly since the development of DiSSCo in recent years. The JRA stream coordinator will be based at RBGE following on from their successful leadership of the JRA in SYNTHESYS3. NHM will act as stream coordinator for the Access programme, having successfully provided the TA helpdesk function in all previous SYNTHESYS projects.

JRA and NA

Each work package will have a leader and a deputy. Work package leaders will host a kick off and subsequent six-monthly teleconference meetings as a means of engaging beneficiaries within their work package. Detailed work plans will be developed and progress will be monitored.

Each task within the work packages has a leader and they will work with the work package leader to ensure deliverables are met on time and within budget. The project manager will attend all kick-off meetings, and subsequent meetings as required.

Work package leaders will be required to report to the Project Council via the JRA/NA stream coordinators, which will act as to correct any deviations and mitigate against any risks identified. The meetings will ensure that links between the NAs and JRA are maintained. The work package leaders will be required to present details of spend so that spend against budget can be monitored. The project manager will ensure that budgets are distributed to match the demands of work being undertaken.

Copies of any internal work package meeting minutes will be sent to the project manager so that she is kept informed of progress. Any problems that may arise within the NAs or JRA will be reported by the relevant work package leader to the project manager who will assist in rectifying the problem, escalating through management chain if required.

Access

Local management of TA and VA is devolved to the TAF leaders, each of whom is supported by a TAF administrator. Their responsibilities include local promotion of the project, answering enquiries, arranging TA user visits and updating the SYNTHESYS database, later ELViS. All TAF leaders are required to report via the Access stream coordinator to the Project Council. The Project Council (including the 13 TAF leaders and the Access stream coordinator) will work to ensure that all TAFs are working to the same transparent standards, so that all users are treated with equity. The Project Council also has responsibility for promoting SYNTHESYS+ TA and VA and ensuring that the users have the best possible experience during their TA visit, enabling new research publications to be delivered.

If a beneficiary's user application rate for TA drops to a level considered unacceptable by the NHMMT, steps will be taken to stimulate further applications for the subsequent call and if that proves unachievable their involvement with TA will cease (the Consortium Agreement will include a mechanism for dealing with such a scenario).

The Access stream coordinator will closely monitor application numbers of the users funded by the TAFs. Revisions to the TAF budgets will take place after each call deadline using the allocation algorithm that was developed in SYNTHESYS2. User days will be reallocated based on demand. This will ensure that the opportunity for each user to be funded is equal, based on demand regardless of the TAF applied to. Gender of applicants

applying and granted access will be monitored by the Access stream leader and the NHMMT and reported at each GM.

3.3 Consortium as a whole

The SYNTHESYS+ consortium represents **one of the largest, and arguably the strongest, consortium of stakeholders associated with NH collections ever constructed**. The incorporation of four major networking bodies (CETAF, GGBN, TDWG and GBIF) adds a level of international reach and expertise that is comparable to the DiSSCo consortium in scope and scale. The group consist of 32 partners including 17 NH museums (including one combined museum/botanic garden), four botanic gardens, four international networks, three commercial partners including two SMEs, two research centres, one university and one not-for-profit foundation. Each adds highly complementary domain expertise representing the breadth and needs of the NH collections community across the nine specific objectives of the proposal (pp.4-5) represented in the nine non-management work packages. Collectively this group bring all the major European NH collections together, alongside a breadth of technical, networking and commercial partners to achieve the overarching aim of SYNTHESYS+ to unify operations and access for European natural science collections. The level of commitment is evidenced by the 25% in-kind contributions provided by all collection-holding partners, and by the exceptionally high number of new partners requesting to join SYNTHESYS (>20 institutions), far surpassing previous requests and the available budget.

Beneficiary 1 has successfully managed the three previous SYNTHESYS contracts and is well-placed to continue in this role. One of the great strengths of the project is that most beneficiaries have worked extensively together as part of H2020 and other projects. All are very experienced working with the NHMMT. The consortium is mature and appropriately structured to deliver SYNTHESYS+, but importantly includes new members, spans new geographic regions and is reaching out to new communities of users.

The SYNTHESYS+ consortium comprises some of the largest and most diverse biological collections in the world, whose holdings and expertise range across the diversity of the natural world, both geological and biological. The long-term experience of managing collections on this scale ideally positions the beneficiaries to carry out the Networking Activity tasks including assaying standards, creating benchmarks and offering advice to collection-holding institutions on how to improve their management, accessibility and long-term preservation. This is complemented by the level of technical competence brought to the consortium by the inclusion of authorities in the domains of AI (A2iA; Digirati), workflow management (UNIMAN) and software development (Picturae).

When combined, the 490 million-strong collections of the beneficiaries are truly global in coverage. When ranked in terms of both relative magnitude and taxonomic diversity European collections are among the best globally. A track record in providing high quality TA can be demonstrated: of more than 3,800 users who completed user feedback reports for SYNTHESYS1-3 96% stated that the collections visited were either excellent or very

good and 89% wished to return to the institution in order to make further use of the collections for their research.

The SYNTHESYS3 Access provision will be extended in SYNTHESYS+ to include **two new TAFs** in Finland (LUOMUS: FI-TAF) and Israel (HUJI: IL-TAF), contributing a combined additional c.20 million specimens, an extended geographic range, and considerable experience and expertise to both the NA and JRA. Existing TAFs DE-TAF and BE-TAF have an additional partner each (ZFMK and BGM), contributing c.10 million specimens and, in the case of ZFMK, an institution focusing only on **animal biodiversity**, thereby complementing the four plant-focused partners (botanic gardens) in the consortium. All four new Access-providing partners have worked with SYNTHESYS previously, and have direct experience of the TA programme through members of their staff carrying out successful TA user projects in SYNTHESYS1-3.

All 21 NH museums and botanic gardens will offer their collections, facilities and staff expertise as part of the TA programme, and 19 will contribute to the new programme of VA. Due to very high demand from European NH collection-holding institutions to join the SYNTHESYS+ consortium, all museums and botanic gardens were required to justify to the NHMMT and SYNTHESYS3 Access steering group why their collections should be included in SYNTHESYS+ and highlight the added value they would bring to both the consortium and more importantly the user community. All were required to complete the SYNTHESYS Collections Self-Assessment, a tool developed in previous SYNTHESYS projects (CSAT), and to have their results audited. This ensured that **all collections offered in the SYNTHESYS+ TA and VA are accessible and under a level of management and care to meet the SYNTHESYS benchmarks**. All TA/VA partners were also required to contribute to the JRA and/or NA, avoiding Access being offered in isolation and ensuring that research and networking activities are integral to achieving the project objectives.

The development and installation of ELViS, the key software supporting SYNTHESYS+ users, is being provided by an **industry supplier (Picturae)** who has considerable experience with several SYNTHESYS+ partners, and a strong relationship with Naturalis who are the lead of this work package (JRA1). In addition, **commercial partners A2iA and Digirati (SMEs)** will represent the domains of artificial intelligence and machine learning, as well as optical character recognition and natural handwriting recognition. Working in close partnership with these companies ensures that the very specific needs of SYNTHESYS+ can be fulfilled by industry and SMEs, who specialise in niche market innovation and the provision of tailored solutions. Working with the diverse SYNTHESYS+ consortium will also provide innovation opportunities for these commercial entities, thus allowing them space to develop new applications and products for use after the project ends.

ELViS activities will include work with the EOSC-hub services to be carried out by members of the EGI Federation, specifically the EGI AAI technology provider GRNET, in collaboration with Picturae. This will ensure synchronisation with the wider European research data and infrastructures community.

All except one of the 32 SYNTHESYS+ partners are automatically eligible to be funded beneficiaries. Beneficiary 32 (Smithsonian-GGBN) is based in the US, will not be requesting any EU funds and will therefore be a partner under Article 9: Beneficiary not receiving EU funding (beneficiary not eligible for funding). GGBN is a key global authority in the field of genomic collections and therefore a natural and necessary work package leader to ensure the success of NA3.

If exploitable IP is generated by SYNTHESYS+, the legal firm Farrer & Co. will be sub-contracted by NHM to provide advice to the originators of the IP on the most appropriate route to market (see 'Management of innovation and IPR' section above). Beneficiaries 2, 9, 16, 23, 24, 26 and 28 will need to use external contractors to undertake Certificate of Financial Statements. These costs are included as direct costs in NA1 (Management) work package. Beneficiaries 1 and 17 have Public Competent Officers that will undertake the certificate of financial statements – there is no cost to the project.

Minor services will be provided by external contractors including printing of promotional material, website support, catering for meetings, leasing of conference rooms and audio-visual equipment, technical support fortraining/promotional video production, and support for promotional events.

Third parties: None planned.

Funding for beneficiaries from third countries: None planned.

Additional beneficiaries: There are a significant number of collections-based institutions that are currently not part of the SYNTHESYS+ consortium which have smaller, but high quality collections that are of widespread interest to potential users for both TA and VA. The previous SYNTHESYS project developed a transparent and open procedure for including collections such as these into the SYNTHESYS+ consortium, in close consultation with the EC Project Officer. This is as follows:

1. Complete a SYNTHESYS collections management self-assessment (to demonstrate the collections are well-managed and accessible) and have results audited.
2. Show demand for Access from European researchers outside of the SYNTHESYS + consortium.
3. When 1 & 2 complete, Project Council to vote on whether to accept the institution as an Access provider.
4. If approved by Project Council, calculate Access costs (including unit cost for TA).
5. Budget allocated as part of the TA/VA budget reallocation to join at the next available Access call.

Beneficiary 17 (MNHN) may seek at a later date to incorporate RECOLNAT (explore.recolnat.org) as a linked third party in order to make the 8.2 million specimens that are part of RECOLNAT available for the TA and VA. To further extend the Access offer internationally, one example of additional beneficiaries are the Komarov Botanical Institute of the Russian Academy of Sciences and the Vavilov Institute of the Russian Academy of Agricultural

Sciences (both in St. Petersburg) who would be strong additions to contribute collections that deal directly with food and crops, thereby contributing to both the SYNTHESYS+ Access programme and the NA. If the demand level of any existing TAF demand level drops to an extent that they are unlikely to be able to meet the user day delivery target, additional partners (including for example the Conservatoire et Jardin botaniques in Switzerland and other key European collection-holding institutions who have strong contributions and are interested in joining the consortium) will be approached. These are the priority institutions on the 'waiting list' of potential partners who have expressed interest in joining SYNTHESYS.

3.4 Resources to be committed (budget)

Overall budget: €11,325,201.90

EU Contribution: €10,000,000

The SYNTHESYS+ budget has been calculated to include significant in-kind contribution from beneficiaries, with 10 million euros requested from the EC. Beneficiaries 18, 20, 25, 30 and 31 (including commercial and SME partners) are claiming 100% costs. Commercial beneficiary 24 is providing 10% in-kind contribution and all other partners are providing 25% in-kind, demonstrating their commitment to the successful delivery of the project.

Management: €703,894 (7%)

Management costs are budgeted at 7% of the total project cost. A substantial cost saving will be made as SYNTHESYS+ will initially utilise the online systems for TA and the consortium management website developed during the previous SYNTHESYS project. Some adaptation will be required to fit with differing EC reporting requirements. Notwithstanding such reprogramming requirements the system is ready for use. The majority of management costs will be contributions towards the NHMMT and stream coordinators' staff costs. The management budget will also include costs for General Meetings, travel and subsistence for dissemination activities, and monitoring visits by NHMMT to User Selection Panels to ensure transparent and consistent practice across all TAFs. The TAFs from previous SYNTHESYS projects now have significant experience in running the TA programme therefore USP monitoring will only be required by teleconference, providing cost savings. TAFs new to SYNTHESYS+, and those with new staff, will be visited in TA call 1 to provide support.

Access: €4,796,996 (48%)

The budget for TA and VA combined is approximately half of the project budget (48%), maintaining a similar portion as for SYNTHESYS3, and allowing the highly successful TA programme to deliver 7,017 user days. However, the Access budget has been split to allow a 20% stake for VA. Both the VA and TA budgets will be reviewed following each call for Access, in order to apportion Access budget to beneficiaries based on actual demand. More budget may be allocated from TA to VA if the demand is proven to be high. Access

costs for the TA include unit costs for all TA partners, together with travel and subsistence costs for TA users and USP meetings. The VA budget is based on estimated actual costs to provide the digitisation activities and associated support and helpdesk functions, largely personnel costs, with a modest (11%) consumables and facilities costs. One beneficiary (2) is claiming subcontracting costs for the VA (maintenance contracts). Significant in-kind contribution is provided by all partners involved in the TA and VA. TAF leaders' and administrators' time providing TA is not charged to the project (either as direct costs or as part of the Access unit costs used by all TA partners). In addition, considerable time input is provided in-kind by all members of the TA User Selection Panel meetings, who score and comment on each application received. In-kind contributions from all VA partners are provided largely in the form of equipment running costs.

Joint Research Activity: €2,343,843 (23%)

The remaining budget is roughly evenly distributed between the NAs (22%, excluding management) and the JRAs (23%). Most JRA costs are for staff time, plus travel and subsistence for work package meetings and specific meetings and workshops to discuss technical requirements. JRA1 includes costs for cloud-based hosting of ELViS, and JRA2 includes consumables and minor equipment costs for tasks 7.2 and 7.3.

Networking Activities: €2,146,188 (22%)

NAs will include largely personnel costs and travel and subsistence for networking meetings, budget for dissemination activities being provided by the management work package. NA5 holds a significantly larger portion for non-personnel costs for supporting stakeholder engagement activities.

Glossary and list of abbreviations

- AAI (Authentication & Authorisation Infrastructure)
- AARC (Authentication & Authorisation for Research and Collaboration)
- ABS (Access and Benefit Sharing)
- ADBC (Advancing Digitisation of BiologicalCollections)
- ALA (Atlas of Living Australia)
- API (Application Programming Interface)
- BHL (Biodiversity Heritage Library)
- BioCASE (Biological Collections Access Service)
- CITES (the Convention on International Trade in Endangered Species
- CONABIO (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad)
- CRIA (Centro de Referência em Informação Ambiental)
- CSAT (Collections Self-Assessment Tool)
- CWL (Common Workflow Language)
- DEST (Distributed European School of Taxonomy)
- DINA (Digital Information System for Natural History Data)
- DiSSCo (Distributed System of Scientific Collections)

- DMP (Data Management Plan)
- DoD (Digitisation on Demand)
- EDIT (European Distributed Institute of Taxonomy)
- EGI (European Grid Infrastructure)
- ELIXIR (European distributed infrastructure for life-science information)
- eLTER (European network of Long-Term Ecosystem Research sites)
- ELViS (European Loans and Visits System)
- EMBRC (European Marine Biological ResearchCentre) (EMBRC-ERIC)
- ENVRI+ (Environmental Research Infrastructures, Horizon 2020 project)
- EOSC (European Open Science Cloud)
- ESFRI (European Strategy Forum on Research Infrastructures)
- EU BON (European Biodiversity ObservationNetwork)
- EUDAT (European Data Infrastructure)
- FAIR (Findable, Accessible, Interoperable, &Re-usable - guiding principles used in relation to data)
- GBIO (Global Biodiversity Information Outlook)
- GFBio (German Federation for Biological Data)
- HTS (High-throughput Screening)
- iBOL (International Barcode of Life Project)
- iDigBio (Integrated Digitised Biocollections)
- IIIF (International Image Interoperability Framework)
- IPBES (Intergovernmental Science-Policy Platformon Biodiversity and Ecosystem Services)
- IPEN (International Plant Exchange Network)
- ISTC (Information Science and Technology Committee of CETAF)
- JRA (Joint Research Activity)
- KPI (Key Performance Indicator)
- MSFD (Marine Strategy Framework Directive, EU)
- NA (networking activities)
- NH (natural history)
- OCR (Optical Character Recognition)
- PRACE (Partnership for Advanced Computing inEurope, Brussels)
- RDA (Research Data Alliance)
- RI (Research Infrastructure)
- e-RIHS (European Research Infrastructure forHeritage Science)
- RRI (Responsible Research and Innovation)
- SANBI (South African National BiodiversityInstitute)
- SME (Small to Medium Enterprise)
- SPECTRUM (an internationally recognisedcollection management standard)
- SPNHC (Society for the Preservation of NaturalHistory Collections)
- TAF (Taxonomic Access Facility)
- TA (Transnational Access)
- TRL (Technology Readiness Level)
- VA (Virtual Access)
- WP (work package)

Members of the Consortium

Summarised in Table 3.

Funding program

SYNTHESYS+ was funded from the RIA - Research and Innovation action in the H2020-EU.1.4.1.2. Programme - Integrating and opening existing national and regional research infrastructures of European interest.

https://cordis.europa.eu/project/rcn/220353/factsheet/en

Grant title

SYNTHESYS PLUS (referred to as SYNTHESYS+ post-submission and in the original description of work text)

Hosting institution

Natural History Museum, United Kingdom

References

- Ang Y, Puniamoorthy J, Pont AC, Bartak M, Blanckenhorn WU, Eberhard WG, Puniamoorthy N, Silva VC, Munari L, Meier R (2013) A plea for digital reference collections and other science-based digitization initiatives in taxonomy: Sepsidnet as exemplar. Systematic Entomology 38: 637-644. https://doi.org/10.1111/syen.12015
- Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2): 81-92. https://doi.org/10.17161/bi.v7i2.3991
- Balke M, Schmidt S, Hausmann A, Toussaint EF, Bergsten J, Buffington M, Häuser CL, Kroupa A, Hagedorn G, Riedel A, Polaszek A, Ubaidillah R, Krogmann L, Zwick A, Fikáček M, Hájek J, Michat MC, Dietrich C, Salle JL, Mantle B, Ng PK, Hobern D (2013) Biodiversity into your hands - A call for a virtual global natural history 'metacollection'. Frontiers in Zoology 10 (1). https://doi.org/10.1186/1742-9994-10-55
- Casino A, Gödderz K, Raes N, Addink W, Koureas D, Hutson A (2018) DiSSCo Partner Capabilities Survey 2017 [Data set - unpublished].
- David B, Taquet P (Eds) (2017) Manifeste du Muséum. [Museum Manifesto]. Reliefs Editions, 80 pp. [In French]. [ISBN 979-1096554263]
- EC Directorate-General for Research and Innovation (2016) She Figures 2015 - Gender in Research and Innovation. Directorate-General for Research and Innovation. Release date:

2016-11-01. URL: https://data.europa.eu/euodp/data/dataset/she-figures-2015-gender-in-research-and-innovation

- EU Council (2014) Conclusions on cultural heritage as a strategic resource for a sustainable Europe, by the Council of the European Union, 20 May 2014. https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/educ/142705.pdf
- European Commission (2017) EN HORIZON 2020 WORK PROGRAMME 2016– 2017 20. General Annexes (Page 29). https://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2016-2017/annexes/h2020-wp1617-annex-ga_en.pdf
- Groom Q, Hyam R, Güntsch A (2017) Stable identifiers for collection specimens. Nature 546 (33). https://doi.org/10.1038/546033d
- Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith V, Triebel D (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017 https://doi.org/10.1093/database/bax003
- Guralnick R, Cellinese N, Deck J, Pyle R, Kunze J, Penev L, Walls R, Hagedorn G, Agosti D, Wieczorek J, Catapano T, Page R (2015) Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. ZooKeys 494: 133-154. https://doi.org/10.3897/zookeys.494.9352
- Hardisty A, Roberts D, Informatics Community TB (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13: 1-23. https://doi.org/10.1186/1472-6785-13-16
- Hobern D, Apostolico A, Arnaud E, Bello JC, Canhos D, Dubois G, Field D, Alonso García E, Hardisty A, Harrison J, Heidorn B, Krishtalka L, Mata E, Page R, Parr C, Price J, Willoughby S (2012) Global Biodiversity Informatics Outlook: Delivering biodiversity knowledge in the information age. Global Biodiversity Information Facility https://doi.org/10.15468/6JXA-YB44
- Hobern D, Baptiste B, Copas K, Guralnick R, Hahn A, van Huis E, Kim E, McGeoch M, Naicker I, Navarro L, Noesgaard D, Price M, Rodrigues A, Schigel D, Sheffield C, Wieczorek J (2019) Connecting data and expertise: a new alliance for biodiversity knowledge. Biodiversity Data Journal 7 https://doi.org/10.3897/bdj.7.e33679
- Kelling S, Hochachka W, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G (2009) Data-intensive Science: A New Paradigm for Biodiversity Studies. BioScience 59 (7): 613-620. https://doi.org/10.1525/bio.2009.59.7.12
- Koureas D (2017) DiSSCo design study summary. https://dissco.eu/sites/default/files/dissco-outline-feb17.pdf
- Kvaček J, Vacek F, Bisang I, Enghoff H, Guiraud M, Haston E, Koureas D, Mergen P, Quaisser C, Smirnova L (2016) D3.6 European Roadmap. http://synthesys3.myspecies.info/node/593
- Purves D, Scharlemann J, Harfoot M, Newbold T, Tittensor D, Hutton J, Emmott S (2013) Time to model all life on Earth. Nature 493: 295-297. https://doi.org/10.1038/493295a
- Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21: 1794-1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x
- Suarez AV, Tsutsui ND (2004) The value of museum collections for research and society. BioScience 54 (1): 66-74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2