Research Ideas and Outcomes 2: e7787
doi: 10.3897/rio.2.e7787

OPEN ACCESS

COST Grant Proposal

Unifying European Biodiversity Informatics (BioUnify)

Dimitrios Koureas[‡], Alex Hardisty[§], Rutger Aldo Vos[|], Donat Agosti[¶], Christos Arvanitidis[#], Peter Bogatencov[¤], Pier Luigi Buttigieg[«], Yde de Jong[»,^], Ferenc Horvath[˅], Georgios Gkoutos[¦,⁇], Quentin John Groom[ς], Tomas Kliment[¢], Urmas Kõljalg[ℓ], Ioannis Manakos[ϑ], Arnald Marcer[P,₳], Karol Marhold[₴], David Morse[₣], Patricia Mergen[₮], Lyubomir Penev[₦], Lars B. Pettersson[₭], Jens-Christian Svenning[₲], Anton van de Putte[⁇], Vincent Stuart Smith[‡]

‡ Natural History Museum, London, United Kingdom
§ Cardiff University, Cardiff, United Kingdom
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ www.plazi.org, Bern, Switzerland
# Hellenic Center for Marine Recearch (HCMR), Heraklion Crete, Greece
¤ Research and Educational Networking Association of Moldova (RENAM), Chişinău, Moldova
« HGF-MPG Group for Deep Sea Ecology and Technology Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremen, Germany
» University of Eastern Finland, Joensuu, Finland
^ University of Amsterdam, Faculty of Science, Amsterdam, Netherlands
˅ Institute of Ecology and Botany, Centre for Ecological Research, Hungarian Academy of Sciences, Vacratot, Hungary
¦ College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, Birmingham, United Kingdom
⁇ University of Aberystwyth, Aberystwyth, United Kingdom
ς Agentschap Plantentuin Meise, Meise, Belgium
¢ University of Zagreb, Faculty of Geodesy, Zagreb, Croatia
ℓ University of Tartu, Tartu, Estonia
ϑ Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
P CREAF, Cerdanyola del Vallès, Spain
₳ Univ Autònoma Barcelona, Cerdanyola del Vallès, Spain
₴ Univerzita Karlova v Praze, Praha 2, Czech Republic
₣ The Open University, Milton Keynes, United Kingdom
₮ Royal Museum for Central Africa, Tervuren, Belgium
₦ Pensoft, Sofia, Bulgaria
₭ Biodiversity Unit, Department of Biology, Lund University, Lund, Sweden
₲ Aarhus University, Aarhus, Denmark
⁇ OD Nature, Royal Belgian Institute for Natural Science, Bruxelles, Belgium

Corresponding author: Dimitrios Koureas (d.koureas@nhm.ac.uk)

Reviewable v1

Received: 14 Jan 2016 | Published: 19 Jan 2016

Citation: Koureas D, Hardisty A, Vos R, Agosti D, Arvanitidis C, Bogatencov P, Buttigieg P, de Jong Y, Horvath F, Gkoutos G, Groom Q, Kliment T, Kõljalg U, Manakos I, Marcer A, Marhold K, Morse D, Mergen P, Penev L, Pettersson L, Svenning J, van de Putte A, Smith V (2016) Unifying European Biodiversity Informatics (BioUnify). Research Ideas and Outcomes 2: e7787. doi: 10.3897/rio.2.e7787

Abstract

In order to preserve the variety of life on Earth, we must understand it better. Biodiversity research is at a pivotal point with research projects generating data at an ever increasing rate. Structuring, aggregating, linking and processing these data in a meaningful way is a major challenge. The systematic application of information management and engineering technologies in the study of biodiversity (biodiversity informatics) help transform data to knowledge. However, concerted action is required to be taken by existing e-infrastructures to develop and adopt common standards, provisions for interoperability and avoid overlapping in functionality. This would result in the unification of the currently fragmented landscape that restricts European biodiversity research from reaching its full potential.

The overarching goal of this COST Action is to coordinate existing research and capacity building efforts, through a bottom-up trans-disciplinary approach, by unifying biodiversity informatics communities across Europe in order to support the long-term vision of modelling biodiversity on earth.

BioUnify will:

1. specify technical requirements, evaluate and improve models for efficient data and workflow storage, sharing and re-use, within and between different biodiversity communities;

2. mobilise taxonomic, ecological, genomic and biomonitoring data generated and curated by natural history collections, research networks and remote sensing sources in Europe;

3. leverage results of ongoing biodiversity informatics projects by identifying and developing functional synergies on individual, group and project level;

4. raise technical awareness and transfer skills between biodiversity researchers and information technologists;

5. formulate a viable roadmap for achieving the long-term goals for European biodiversity informatics, which ensures alignment with global activities and translates into efficient biodiversity policy.

Keywords

COST Action, Biodiversity Informatics, Environment, Standards, Data interoperability

Context

In this article we publish the full text of the proposal for a new COST Action titled "Unifying European Biodiversity Informatics (BioUnify)". The proposal is presented as submitted on 10 April 2014 to the open call of the COST Association with a reference ID OC-2014-1-18556. As with all submissions to the COST Association calls, the proposal was evaluated without revealing to the reviewers the identity or affiliations of the authors, or the network of supporting organisations. A total of 65 organisations from 24 countries supported BioUnify (41 higher education & associated organisations, 15 other government/ intergovernmental organisations, five private non-profit organisations and four business enterprises).

Proposal rationale

Biodiversity is the study of the diversity of life at all possible levels of the biological organisation (from genes to ecosystems) and scales of observation (from local to global). Therefore, studies of biodiversity are predicated on the capacity to bring together information from across a diverse spectrum of scientific fields. For more than a decade and as the volume of available information is increasing several projects were initiated and organisations focused on better organising this information. Significant steps were made in developing tools and services for mobilising and aggregating biodiversity related data at regional and global scale. Alongside, community standards were developed to facilitate data and system interoperability.

Given the scale and urgency of the societal challenges related to environment, better coordinated efforts are required to enable the linking of diverse datasets and provide unified and easy to use services to multiple audiences. Researchers, policy makers and the public are all in need of seamless access to services that enable access to and reasoning with diverse and complex datasets. BioUnify brings together key players from the biodiversity and informatics communities, for the first time at this scale, to create a umbrella structure that enables scientific and technological innovation. Innovation focused on achieving harmonisation of tools, services and datasets.

BioUnify aims at developing a common technical backbone, addressing issues related to data quality and fitness-for-purpose, and enhancing data skillsets for scientists. It builds on existing efforts, providing a path to harmonisation and unification of initiatives across Europe, in response to biodiversity and wider environmental challenges.

Following the unsuccessful submission of BioUnify, the community dispersed efforts towards supporting interoperability through existing cross-domain collaboration and networking platforms. For instance through the Research Data Alliance - Biodiversity Data Integration Interest Group, and other Interest and Working Groups. The overarching challenges, however, still stand. Coordination actions at a global scale are still needed to support scientific and technological research as well as develop user friendly services underpinning the community's long-term vision of modelling all life on earth.

Evaluation outcome

Following evaluation by three external reviewers the proposal received an average mark of 29.33/40 and was subsequently not selected to be funded by the COST Association.

The consortium received the reviewers' comments, which we summarise in Table 1. The key points are presented for each of the evaluation sections: State of the Art, Relevance and Timeliness, Feasibility, Risk level, Scientific and/or Societal Impact, and Timeframe. The authors did not receive permission, from the COST Association, to publish the full text of the reviews.

Table 1.

Summary of the key points from the anonymous reviews of the BioUnify proposal. The points presented are a high level interpretation, by the authors, of the evaluation feedback received and include both positive aspects and shortcomings. (+) Positive comment, (-) Identified shortcoming.

Evaluation question	Key points
Understanding of the State of the Art	• (+) A good understanding of the state-of-the-art with very clear and updated contextualisation of biodiversity research through informatics; • (-) Not clearly addressing the overarching scientific questions relevant to the technological challenges presented; • (-) Not sufficiently recognising the role of key stakeholders, such as the management bodies of protected European sites and environmental agencies.
Relevance and timeliness of the proposal	• (+) Clearly identified, highly relevant challenge; • (+) Timing pertinent as problems still persist; • (-) An earlier start of the proposal might have been more impactful.
Challenge feasibility	• (+) Feasible challenge; • (+) Multi-disciplinary approach very positive; • (-) Slightly overambitious; • (-) Lack of evidence on its ability to influence decisions on global or European scale.
Risk level	• (+) Well-established proposal, with outcomes presented in confident manner; • (+) Proposal return of high potential impact; • (-) Uncertainty of success given the native complexity of biodiversity data; • (-) Doubts on the availability of required datasets, methods and algorithms for the purposes of the Action.
Scientific and/or Societal Impact	• (+) Strong and motivated network; • (+) Success of the Action would positively imapct scientific community, providing common biodiversity standards and enabling interoperability; • (-) Sparse evidence on ability to provide firm impact; • (-) The path to solve issues was not adequately explained.

Timeframe	• (+) Clear structure of the proposal and large consortium; • (+) Convincing in demonstrating the need for networking to meet the challenge; • (+) Anticipated results in short-, mid- and long-term; • (-) Lack of clear guiding conceptual framework of scientific relevance for the network composition and operation method; • (-) Timings over-ambitious; • (-) Innovation rather limited due to lack of conceptual framework that shows the practical development guidelines as the solution.

Challenge

The need for biodiversity Information management

The strategic plan for Biodiversity for 2020, including the Aichi Targets, has prioritised the need for effective study of biodiversity at a global scale. As a result, understanding and protecting biodiversity has become the main pillar for tackling many of the societal challenges both at a regional and global level. Policy makers are urgently in need of means to monitor the status and trends of life on Earth, predict the impact of changes, and support the right policies to minimise the depletion of the planet's biological diversity. In the race towards this long-term goal, the cornerstone action is to effectively bring together data and information in a way that enables researchers to establish correlations, identify patterns and produce knowledge that yields novel insights or explanations.

Access to more efficient and affordable research infrastructures, including powerful computing facilities, innovative environmental sensors and other instruments, is increasing the volume of data at an unprecedented rate. From next generation sequencing to remote bio-sensing methods and natural history collections digitisation programs, data are flooding the scientific community. Sometimes the uses of these data are not immediately obvious or lack provisions for interoperability. As the volume increases, fundamental questions arise on the ability to use and reuse these data in a way that enables researchers to understand ecological and evolutionary processes and model complex biological systems. How do we curate, preserve and process the increasing volume of biodiversity data? How do we identify gaps in available data or re-combine existing datasets for use across different biodiversity disciplines? And finally, how do we facilitate the entire data lifecycle from generation to publication and reuse?

The systematic application of information management and engineering technologies in the study of biodiversity (biodiversity informatics) provides methods and tools that can facilitate the entire big data lifecycle and eventually help transform data to knowledge. Biodiversity informatics is concerned with improving the management of data, information and knowledge from molecular to ecosystems level, and supports a more holistic approach to the study of biodiversity. In this regard, informatics tools and services are crucial to systematically and reliably assess global biodiversity changes and make coherent and robust predictions about ecosystems.

To achieve a high-level of integration of information technology (IT) in biodiversity research it is fundamental to develop and sustain a network of skilled people (data scientists) across different biodiversity research disciplines and nurture effective synergies between researchers, software engineers and information technologists. Only by facilitating a productive communication cycle between these scientific and technological domains will biodiversity informatics provide viable and sustainable solutions that address the societal grand challenges that result from global change. This requires new ways of doing research that better link science and society to address the needs of decision-makers and citizens at global, regional, national, and local scales.

Biodiversity informatics underpins the data and processes needed to develop models that will predict biodiversity, ecosystem degradation and regeneration rates. These models can translate into a sustainable biodiversity and ecosystem utilisation policy that promotes efficient conservation and agricultural practices. Biodiversity research needs to invest in a whole-system, synthetic approach, with semantic interoperability across all taxonomic, genomic, ecological, agricultural and marine data; based on a small set of interchange standards. This can only be delivered by fostering a culture of interdisciplinarity and collaboration. A trans-domain approach is required to bring about the level of efficient data mobilisation, information management, and analytical platforms necessary to achieve this vision.

Current state-of-art

The need for efficient informatics tools in biodiversity research is constantly increasing. This statement can be supported by the volume of different biodiversity information projects (>680) (http://www.tdwg.org/biodiv-projects/) currently running at a local, regional or global level. Research and data management organisations across Europe, including academic institutions and natural history collections, have invested vast resources in the de-novo development of tools to support in-house data management and processing. It is more often than not that these custom-made tools are duplicating previous efforts, lack provision for open access or semantic interoperability, and are inherently over-specialised.

Biodiversity informatics originally focused on developing specifically adapted technological solutions in response to the diverse nature of data types produced by the different communities. This somewhat framentary approach took precedent over the longer-term need to devise and implement domain-wide standards for data, interfaces and processes. These parallel and investigative activities, however, inspired exploration of alternate approaches and underpinned innovation. Over the last decade, biodiversity informatics research has reached a level of maturity that requires assessment and consolidation of applied technologies through coordinating efforts. Efficient pan-European and worldwide collaboration will consolidate and harmonise the biodiversity informatics landscape, leveraging the ability of invested resources to deliver results within the scope of big scientific and societal challenges.

On a European level, numerous biodiversity informatics projects have been funded by the recursive European funding mechanisms. Networks of excellence, including ALTER-Net, LTER-Europe, EDIT/PESI, MARBEF/EuroMarine along with other projects including 4D4Life/i4ife, agInfra, AquaMaps, iMarine, BioFresh, BioVeL, ENVRI, EU-BON, EU-BrazilOpenBio, Fauna Iberica, MicroB3, OpenUp!, pro-iBiosphere, BioSOS and ViBRANT, have created a portfolio of research and e-infrastructure approaches for biodiversity data. Most of these projects have acknowledged and addressed, within their respective lifetime, the need for open access, data interoperability and community capacity building. Nevertheless, the information and system architecture, and technological approaches used, differ substantially between each project, demonstrating the lack of an agreed upon technical and procedural model for delivering interconnecting services. Individual projects have addressed these issues by introducing methods for enhancing syntactic (data formats) and semantic (meaning of data elements and relationships between them) interoperability, by highlighting the need for pan-European coordination of activities and by addressing the gap between biodiversity researchers and information technologists.

As part of the the European Strategy Forum on Research Infrastructures (ESFRI), initiatives like LifeWatch have started. LifeWatch aims to provide a biodiversity and ecosystem research infrastructure based on a coherent Europe-wide (top-down) plan, and by encouraging implementation of solutions locally (bottom-up). LifeWatch could eventually deliver efficient coordination activities, but does not yet have the necessary mechanisms in place to underpin the bottom-up coordination, which is so urgently needed to leverage biodiversity and biodiversity informatics research in Europe today. This COST Action (BioUnify) will act to initiate a viable rapid response collaboration platform to act as this bottom-up coordination, organising the community under the umbrella of community-defined scientific and technical objectives. It will work with the top-down thinking, providing the necessary effort to effectively translate strategy into clear guidelines for biodiversity data owners, data custodians, tool developers and researchers.

Several global initiatives have been established for enhancing biodiversity research through informatics. These include: Catalogue of Life (CoL), Biodiversity Information Standards (TDWG), the Global Biodiversity Informatics Facility (GBIF), Encyclopedia of Life (EOL), the Biodiversity Heritage Library (BHL), the World Register of Marine Species (WoRMS), the Ocean Biogeographic Information System (OBIS) and the Genomic Standards Consortium (GSC). Projects like CReATIVE-B and GEO BON address convergence through encouraging international liaisons and common activity. Biodiversity informatics research, as defined by the objectives of all these projects, is a highly inclusive trans-disciplinary domain, able to demonstrate significant research results in the areas of:

- Development of biodiversity data exchange standards;
- Mark-up activities and mobilisation of data from biodiversity legacy literature and natural history collections;
- Virtual research environments and mobilisation of long-tail data;
- Semantic interoperability and domain specific knowledge organisation systems;
- Cloud based computational tools, analytical services and application workflows;
- Data cleaning and data harmonisation.

As research in all these areas advances, it is crucial to foster reciprocal interactions between information engineering and biodiversity research, in a timely and efficient manner. Close interactions between individuals with skills in these domains are essential. Despite previous initiatives, the biodiversity landscape in Europe is still characterised by a high-level of fragmentation, with minimum functional interactions and with prominent elements of duplication. The lack of a bottom-up, self-driven organisation and coordination platform is hindering the two wider communities from joining forces. This issue has been repeatedly reported in scientific conferences (e.g. Biodiversity Informatics Horizons 2013, Rome), scientific publications (e.g. Hardisty et al. 2013) and published reports (e.g. Hobern et al. 2013). The challenges described below are defined by the need for translation of biodiversity research goals into clear technical specifications for the development of robust technical solutions, raising awareness and efficient training.

The challenges

The biodiversity informatics domain today faces a series of technical, sociological and decision making challenges; from adopting an industry-standard technical backbone that underpins its activities, to effectively bringing together data and communities across scientific domains and between different disciplines within these domains.

The urgent need for information solutions has been the driving force for developing existing infrastructures but repeatedly inventing bespoke solutions, which often feature overlapping elements, is the norm. European biodiversity informatics is today at a pivotal point where a pan-European (if not global) approach is needed to unify biodiversity informatics and deliver the level of services necessary for biodiversity research to reach its full potential. Biodiversity informatics research is funded both at national and European level. The result is the deployment of a series of unrelated e-infrastructures, platforms and software. To make good use of the research efforts invested, it is critical to: (i) enable interactions that will produce synergies between software engineers, data architects and data custodians in order to identify components that can be effectively brought together as part of a global solution; (ii) assess the existing landscape and propose a set of models and services to be used at an industry-wide level, reducing future proliferation, and (iii) raise awareness to end users and key stakeholders by producing documents with potential impact on policy making processes in Europe. Together these actions will help to converge institutional and national thinking at a European level.

A common technical backbone

At the core of biodiversity informatics is research for developing and applying technological solutions that facilitate biodiversity data management, structuring, linking, analysis, visualisation and reuse. Over the last decade, several technical models have been developed and sporadically applied in European e-infrastructures. This set of models, however, has not been systematically assessed for its interchangeability and its effectiveness in mobilising data across the entire spectrum of biodiversity research.

Arguably, the challenge of delivering robust and predictive models for biodiversity in Europe and worldwide, is dependent on the ability to create links between existing and newly generated data across the biodiversity domain and related domains (e.g. that of climate data). Effective collaboration between biologists and informaticians is essential to accomplish this aim through identifying, modelling, and quantifying logical connections between heterogeneous datasets in a systematic and sustainable manner.

A set of common biodiversity data standards are crucial for achieving interoperability. These facilitate storing and exchanging both data and computational models between machines. Within the biodiversity research and e-infrastructure domains, a mixture of both generic and domain-specific standards have been used. Bespoke standards like the Taxonomic Concept Schema (TCS), Access to Biological Collections Data (ABCD) and the Structured Descriptive Data (SDD), provide more structured ways of storing taxonomic, phylogenetic and related biodiversity data, while others, like the Darwin Core (DwC) are based on more simplified schemas. Related standards like DwC-A, EML or DublinCore are used for machine to machine data/metadata exchange. Extensions for most of these data schemas have been developed to handle heterogeneous data types (i.e. taxonomic concepts, specimen records, ecological field records, phylogenetic data, sequencing metadata and media metadata). This work is being extended into new areas such as environmental and biodiversity genomics, linking Darwin Core with the MIGS/MIMARKS standards from the Genomics Standards Consortium. Nevertheless, the biodiversity informatics community will need to assess these standards via a more systematic approach to deliver a common set of simple interchangeable, scalable and extensible data schemas with ability to mobilise broad interdisciplinary data types. Issues related to the levels of granularity of the data fields, the use of effective metadata, and the widespread use of common data formats across biodiversity e-infrastructures, are key milestones for developing a common technical backbone.

The foremost important element in performing interdisciplinary research drawing from the big data pool is to be able to work with accurately annotated and described data. Well structured, semantically enriched datasets can be interlinked in such way that large-scale reasoning can be performed. The most important aspect of this approach is the ability to use efficient and comprehensive controlled vocabularies and ontologies to annotate and structure data. Comprehensive taxonomic checklists of animals, plants, fungi and microorganisms as well as structured bio-ontologies (including environmental ontologies) are essential. Minimising existing redundant elements and improving the inclusiveness of ontologies will be critical in leveraging their application and use. This requires ontologists to work closely with domain specific research scientists that generate or curate biodiversity data.

One of the main requirements for efficiently linking biodiversity data from distributed databases is the assignment of global, unique and stable identifiers. Identifiers enable networked services to locate and link to different resources. For certain types of resources, including scientific publications, significant progress has been made in globally applying unique identifiers (i.e. DOIs), while for others, including published biodiversity datasets, specimen data, taxonomic concepts, computational workflows or even environmental

samples, the community has still not adopted a common model of unique and persistent object identification. A vibrant discussion is currently sustained at a global level, on the preferable model and implementation method of unique identifiers. UUIDs, Handles, DOIs, PURLs, HTML URIs and LSIDs are some of the systems currently used by biodiversity information systems for uniquely identifying digital objects. Each with its own merits and shortcomings, the European biodiversity informatics landscape will need to consider adopting a common model for unique and actionable, i.e. resolvable, identifiers. A bigger challenge, however, will be to advocate the need for using persistent identifiers in all the biodiversity related databases that are exposed and accessible online. This will require coordinated efforts for persuading database curators and institutions on the added value of implementing these systems.

On the top of the technical pyramid that underpins biodiversity data, is software that delivers useful, efficient and simple-to-use functionality to scientists. Software engineers and informaticians need to sustain active communication channels with biodiversity researchers in order to design, develop and deploy software that supports research activities and incentivises researchers to make use of the advantages that online, collaborative and open science brings.

Data mobilisation and fitness-for-purpose

Biodiversity-related data are being generated and/or managed by a wide spectrum of stakeholders including individual researchers, research groups, governmental and intergovernmental agencies and natural history museums. These data usually come in native formats, often with minimal provision for standardisation or long-term accessibility, and without thorough documentation.

Biodiversity research is experiencing a bloom of new data, but the rate of increase is not equal for all the disciplines. As the big-data pool deepens, it is becoming more difficult to identify data types that fall behind. Genomic, biosensing, taxonomic and natural history collection data are produced at ever increasing rates. Data aggregation initiatives are more successful for some data types (e.g. GenBank, GBIF, EOL, BHL), while for other data types (e.g. ecological traits, morphological characters, field observations) progress is slower and more complex (e.g. EMODNET biology, TRY-DB, LifeWatch, EOL TraitBank, iPlant Collaborative). Given these discrepancies, it is crucial to assess data generation rates across these disciplines, in the context of the importance of the data to other biodiversity communities, and in the scope of a long-term vision to create viable models of biodiversity with robust predictive power.

Short-term storage and long-term archival solutions for the heterogeneous and large-volume data produced across the different biodiversity disciplines is already a major problem. This is vital to facilitate horizontal knowledge transfer across stakeholders and to foster the development of sustainable initiatives for aggregating, archiving and linking biodiversity data. The inconsistent and limited application of global unique identifiers for data entities exposed on the web makes this an especially difficult challenge. Only by addressing this problem, will it be possible to sustainably use major biodiversity data

infrastructures to provide reliable web-services that retain the level of granularity needed to secure data fitness for diverse purposes.

Education, training and awareness

Biodiversity research is rapidly migrating to the digital domain. Digital data aggregators, repositories and registers are now providing significant new opportunities for data extraction, interlinking and re-use. These e-infrastructures are invaluable sources of information for researchers around the world. Despite their value, these new tools are characterised by low uptake rates within many research communities. To facilitate the digital lifecycle of biodiversity data, from generation to publication and re-use, it is vital that biodiversity researchers understand the opportunities and limitations of these tools and embed them in their day-to-day research activities. By training a new generation of IT-literate biologists and biology-aware computer scientists, biodiversity informatics will deliver the knowledge needed to meet demanding societal challenges in Europe and worldwide. A combination of ad-hoc training activities, mentoring programmes and an established graduate curricula, are needed to produce a new generation of Biodiversity data scientists. These individuals need to have the analytical and data curation skills required to understand, handle and process big and heterogeneous datasets.

Added Value of Networking

This Action (BioUnify) draws from the experience that biodiversity cannot be fully understood by the work of any individual discipline, but through an integrative approach of cross-disciplinary research. Effective coordination is required to synthesise available data in order to develop robust analytical and predictive models and communicate the output of these models through informative visualisations, helping shift research priorities. BioUnify will enable efficient interactions between people from different biodiversity and informatics disciplines under the umbrella of biodiversity informatics. How can effective integration of taxonomic, genomic and ecosystem research be achieved? How can diverse datasets plug into harmonised research workflows to test application-level biodiversity models? How can data interoperability act as the platform for biodiversity interdisciplinary research? These questions can only be answered through operating and sustaining an extended, but also highly structured, network of stakeholders that promotes synergies across expertise and supports cross-fertilisation of ideas.

Scientific dialogue is naturally achieved through established forms of scholarly communication, including scientific publications and conferences. These channels are time consuming and do not necessarily focus on addressing specific and urgent technical or societal issues. Agile and effective communication between people, at the level (across scientific domains and communities) and timeframe needed to address explicit societal challenges, demands a highly focused network of people and activities. A network that will enable researchers to jointly shape research goals and adjust methodologies for delivering results in scope and on time.

This COST Action will address the challenges described above through coordinating the biodiversity informatics community and by creating a wide trans-disciplinary collaboration platform. This will improve the efficiency of distributed funded research activities to enhance collaboration of isolated research groups, under an umbrella of commonly defined scientific goals.

BioUnify will:

- Assess the existing technical specifications for mobilisation and linking of heterogeneous biodiversity research and monitoring data;
- Investigate possible solutions by adopting a high-level and integrative approach and propose amendments/changes to existing models increasing their interchangeability and inclusiveness;
- Foster technological innovation by nurturing new collaborations between research scientists and technologists towards the long-term vision of delivering robust predictive biodiversity models;
- Monitor the course of actions of relevant international organisations and position its activities within their wider strategic framework;
- Improve the influence of the European biodiversity informatics community over decision making processes internationally, by sustaining efficient communication channels with global actors and stakeholders, and by promoting the authoritative role of European institutions;
- Propose a simple, open access, and interoperable set of models that stakeholders, generating or curating biodiversity data, could embed in their day-to-day workflows; and finally
- Bridge the existing gap between biology researchers and informatics technicians by combining the expertise and promoting the training and the required capacity building.

Building upon existing efforts

Data and e-infrastructures are capital investments, that are very expensive to regenerate or rebuild. This COST Action will build on existing efforts developing a coordination network that brings together biodiversity researchers and software engineers (including information technologists) to maximise the efficiency of already running projects and create a reference platform for the development of future activities.

Researchers already involved in biodiversity informatics projects have acquired unique domain experience. Similar experience has been accrued by researchers that have been making use of related e-infrastructures. Their indispensable input, along with knowledge extracted from key scientific publications and international reports, will be codified to form the basis of knowledge and capacity for this Action.

Supporting biodiversity research

Several key issues have been identified in the literature that hinder cross-disciplinary biodiversity research; Lack of persistent identifiers in all digital produced objects; lack of appropriate registers/aggregators; the absence of industry-level knowledge organisation systems that enable semantic interoperability and data reasoning; an absence of a common taxonomic backbone, essential for linking dispersed information; and finally the large volume of “dark data” (inaccessible legacy data) produced through long-tail research (P. Bryan Heidorn 2008). These issues significantly slow down, if not restrict, interdisciplinary and big-data research on biodiversity. BioUnify will translate the results of its activities, through its four-year lifespan, into clear recommendations and will use its extensive network of people as ambassadors to their local and discipline-specific communities.

Coordinating key stakeholders

Individual researchers, research communities, biodiversity organisations, natural history museums, policy and decision making instruments, approach the study of biodiversity from different points of view. The need for effective communication and coordination across the entire spectrum of these stakeholders has long been acknowledged by the community. Fragmentary efforts, within certain projects and/or discipline-oriented networks, has already initiated a dialogue between some of these stakeholders. This Action will codify the research and policy priorities of different stakeholders of the biodiversity and biodiversity informatics domains and translate them into actions for achieving convergence between biodiversity data, modelling protocols, information and knowledge.

Developing a roadmap for the future

The biodiversity informatics community has set its long-term vision of supporting people, data and processes to deliver robust models of the biosphere and to apply the predictive power of these models in addressing societal challenges. Inevitably this vision goes beyond the lifetime of this COST Action. BioUnify will develop the mechanism to nurture collaboration and research coordination, but will also develop a detailed roadmap for reaching jointly set long-term goals. This roadmap will be focused on describing and time-framing the required future steps. It will be formulated as a series of best-practice documents that will be widely disseminated through multiple communication channels. Such a roadmap can act as an umbrella for positioning future research proposals.

Providing training and raising awareness

Training early stage researchers in making use of available technologies in their respective research domain should be considered crucial. Advances in information management and newly developed tools can only be established through wide adoption by the majority of researchers, and in the context of open access research activities. Proper training will ensure that researchers are aware of the available tools and services, make efficient use of

them, and develop their research careers in the context of long-term visions linked to societal issues and cutting edge research.

BioUnify will adopt a hybrid model for providing training. Direct training will be given by organising workshops and supporting Short Term Scientific Missions (STSM) for both Early Stage (ESR) and Experienced Researchers (ER). Emphasis will be given to supporting ESR mobility between groups with different expertise. This will provide ESRs with diverse skillsets that combine biodiversity research with IT and software engineering competencies.

Organised training courses, workshops and hackathons will bring together a heterogenous set of people that spread across the different scientific domains and disciplines that this Action is relevant to. This will initiate useful interactions between all participants in the context of promoting interdisciplinary and trans-domain research. The second pillar of this hybrid model is to make effective use of all established training mechanisms, including higher education training, by providing reference material and human expertise (e.g. invited lectureships) to postgraduate courses on biodiversity and biodiversity informatics. Furthermore, BioUnify will support initiatives to embed biodiversity informatics training in related postgraduate programmes running by universities in Europe.

Provisions for post-Action sustainability

Fostering interdisciplinary research goes beyond breaking technical barriers that restrict the mobilisation and sharing of data and information: This is also a major social/cultural issue. This Action will balance providing technical solutions through biodiversity informatics tools with a culture change programme that nurtures an open access and trans-disciplinary research philosophy.

Through targeted Short Term Scientific Missions (STSMs) the Action will encourage researchers to test innovative ideas bridging different scientific disciplines and domains. This Action will enhance collaboration and promote the necessary cross-domain synergies, but will also focus on supporting the development of pan-European initiatives that will seek further funding to support the Action's objectives and to sustain the Action's activities beyond its lifetime. BioUnify will identify calls and related topics from which funding can be secured (e.g., national sources, Horizon 2020) to continue supporting the delivery of robust information solutions for the biodiversity research domain. Towards this end the MC meetings will be used to form the networks necessary to deliver new consortia. Finally, the Action will lead to the creation of a permanent entity that will continue horizon scanning, scoping for new opportunities for interactions between biodiversity researchers and information technology specialists. The new entity will consist of European academic and research institutions, natural history museums, scientific societies, biodiversity related governmental and intergovernmental organisations and SMEs. During the Action, sustainability issues, both from a financial and a societal perspective will be addressed. Long-term sustainability issues of the new entity will be considered.

Complementarity to existing projects/networks

Interactions at a national or European level

As the overall vision of this Action is to leverage existing efforts it is crucial to develop a collaborative platform that will take into account all related projects and initiatives, create active links and provide useful feedback.

This COST Action will be highly complementary to existing networking or projects that are running within the biodiversity informatics domain. Networking activities are included in most of the biodiversity e-infrastructure projects funded at a European level. These activities, however, often only include the limited number of partners. BioUnify will build upon these networks, enhance them and expand them. As a multi-disciplinary Action, BioUnify objectives align with and complement the objectives included in other COST Actions (e.g. ESSEM: HarmBio, ESSEM: EMBOS, ICT: KEYSTONE, FA: ALIEN Challenge) and ongoing FP7 projects (e.g. EU-BON). Furthermore, BioUnify will cross-fertilise other national or regional biodiversity and biodiversity informatics research projects and act as a proxy, facilitating horizontal transfer of knowledge and skills. Provisions will be made to coordinate the Action's activities in alignment with the Biodiversity Information System in Europe (BISE) and the Biodiversity Data Centre (BDC). The Action will act as an essential stepping stone towards the development of an operational network of biodiversity research and e-infrastructure networks in Europe and directly contribute to the overarching scope of related ESFRIs (e.g. LifeWatch).

Interactions at an international level

BioUnify aims at contributing to the scope of the European Research Area by initiating communication with peer networks across the world (e.g. USA Resource Coordination Networks) and present a unified pan-European voice of authority. The challenges described in section A go beyond the European boundaries. As such, capacity building efforts in Europe will need to be invested against and compared to the global landscape. Creating strong links with the newly initiated Intergovernmental Platform on Biodiversity & Ecosystem Services (IPBES) will be fundamental. This Action will deliver its outputs in the framework of activities by European consortia, as CETAF and its group of interest on bioinformatics (ISTC), and by global organisations, including the Research Data Alliance (RDA) and the Biodiversity Information Standards (TDWG). Where applicable, the members of the Action will actively participate in associated Interest Groups and Working Groups of these organisations. In this regard the Action will build upon the output of previously funded international cooperation projects on biodiversity research (e.g. EU-Brazil OpenBio, CReATIVE-B, COOPEUS).

Milestones and Deliverables: contents and time frames

Action outputs

The objectives of the Action are defined against the challenges described in section A. BioUnify has four objectives that fall under respective COST Objective categories. The MC will be able to amend and adjust the list of deliverables to increase the efficiency of the Action and align its activities with the perpetually changing biodiversity informatics landscape.

Objective 1: Models and systems assessment

Task 1. Data storing and exchange schemas

Within this task the current biodiversity data schemas, used to mobilise data, will be assessed and technical shortcomings will be identified. This will include data exchange formats used to communicate structured data between data providers, aggregators and registers and between data management platforms and publishers. The assessment will be performed in the context of harmonising existing solutions and supporting the mobilisation of diversified data types. Assess, in terms of feasibility and fitness-for-purpose currently used common inter-linked identifiers for different data entities, across information systems at web-level.

Deliverables/Milestones: 4 Short Term Scientific Missions (STSM) (Yr1, Yr2, Yr3, Yr4); 3 open access peer-reviewed publications (OA-P) (Yr2, Yr3, Yr4); Workshop/Hackathon (WS/H) (Yr1); Joined Student Supervision (JSS) (Yr1)

Task 2. Controlled vocabularies and ontologies

Catalogue and deliver fitness-for-purpose reports on the available knowledge organisation systems (including controlled vocabularies and ontologies). Support the further development of existing bio-ontologies as the mean for creating logical connections between heterogeneous biodiversity data. Investigate the wide application of (among others) Global Names Architecture and the Environmental Ontology (ENVO) as focal points for linking biodiversity related data.

Deliverables/Milestones: 4 STSMs (Yr1, Yr2, Yr3, Yr4); JSS (Yr2); WS/H (Y2); 3 OA-Ps (Yr2, Yr3, Yr4)

Task 3. Service networks, computational workflows and virtual research environments

Evaluate existing solutions for building and sharing computational workflows and assess the bottlenecks in integrating, within these workflows, existing data provider services and processing/analytical tool services. This action will identify gaps in the service network provision of virtual research environments and an economic cost evaluation of the

computing resources necessary to sustain these activities. Data longevity and services sustainability will be taken into consideration across activities under this Task.

Deliverables/Milestones: 4 STSMs (Yr1, Yr2, Yr3, Yr4); JSS (Yr3); WS/H (Yr3)

Objective 2: Data availability, interoperability, harmonisation and long-term preservation

Task 1. Natural history collections data

Summarise current global practices and workflows followed by major natural history museums, natural science museums and collection holders (individually and using aggregate data provided by existing consortia), for specimen digitisation, specimen data management and metadata extraction and exposure. Evaluate existing database models used and archiving solutions implemented.

Deliverables/Milestones: 2 STSMs (Yr1, Yr2,); WS/H (Yr1); OA-P (Yr1); Consolidated Report (CR) (Yr1)

Task 2. Biodiversity, genomic and ecosystem research and monitoring data

Bring together information technologists and leading research groups in Europe that generate or curate big biodiversity, genomic or ecosystem datasets. Identify commonalities in data exchange models used to store and mobilise data within these communities. Evaluate fitness-for-purpose of curated databases and identify critical elements in data models used, including data provenance and data versioning.

Deliverables/Milestones: 4 STSMs (Yr1, Yr2, Yr3, Yr4); JSS (Yr2); WS/H (Yr2); 3 OA-Ps (Yr2, Yr3, Yr4)

Task 3. Biodiversity legacy data mobilisation

Investigate the technological solutions available for extracting, annotating, mobilising and digitally preserving data from legacy sources, including biodiversity legacy literature. This task will aim at summarising existing approaches and workflows, consolidate outputs from individual projects that tackled this issue and identify common technical challenges. Bring together legacy data custodians, publishers and experts in mark-up activities.

Deliverables/Milestones: 4 STSMs (Yr1, Yr2, Yr3, Yr4); JSS (Yr2); WS/H (Yr2); 3 OA-Ps (Yr2, Yr3, Yr4)

Task 4. Data publication, long-term preservation and re-use

Data publication, long-term preservation and data re-use are critical properties of biodiversity information management, as they constitute the primary incentives that motivate key actors (including individual researchers) to embed information solutions in their day-to-day research and data curation activities. Publishers, registers/aggregators as well as archiving infrastructures need to act in a highly coordinated way to facilitate the

seamless data cycle. This Task aims at summarising existing publication, aggregation and archiving workflows used and coordinate ongoing activities to enhance existing or develop new more efficient innovative workflows.

Deliverables/Milestones: 4 STSMs (Yr1, Yr2, Yr3, Yr4); JSS (Yr3); WS/H (Yr3); 3 OA-Ps (Yr2; Yr3; Yr4)

Task 5. Action Conferences

The Action Conferences will be organised on the broadest possible trans-disciplinary approach and aim at bringing together biodiversity data providers, managers and consumers in an attempt to enable viable interactions and describe minimum requirements for generating and curating interoperable datasets. Breakout groups will specifically aim at enhancing collaboration between information and biodiversity scientists. Conferences will be supported by the COST Action as well as by associated nationally and European funded projects.

Deliverables/Milestones: 2 Conferences (Yr1, Yr4); 2 Conference Proceedings (Y1, Yr4)

Objective 3: Education, Training and Dissemination

Task 1. Researcher training and awareness

Arguably, delivering training for skills development and raising awareness on the impact and value of deploying efficient information solutions to manage, share and re-use biodiversity data is the cornerstone for effectively improving biodiversity informatics uptake in the long-term. This Task will focus on identifying the primary skills that new researchers will need to have and identify the channels through which these skills can be transferred. Furthermore, will organise and deliver high-quality training through a series of training schools aiming at Early Stage Researchers.

Deliverables/Milestones: 4 Training Schools (Y1, Yr2, Y3, Yr4); Workshop (WS) (Yr1); CR (Yr1)

Task 2. Biodiversity informatics training modules

Following a hybrid model of delivering training, this Task will catalogue already developed training modules (e.g. CETAF, DEST, SYNTHESYS, ViBRANT) that aim at transferring specific informatics skills to biodiversity researchers and summarise available diplomas related to biodiversity informatics in existing higher education structures. The impact of those modules will be evaluated and recommendations will be made to enhance their efficiency. Furthermore, will facilitate, through the website, access to existing training material and announce related training courses in Europe and internationally.

Deliverables/Milestones: WS (Yr2); CR (Yr2)

Task 3. Action website

The Action website will act as a focal point of all the activities. The website will be build using existing available open-access and free to use infrastructure and will be used as an online project management platform, as a dissemination mechanism and source of project outputs. Provisions will be made in order to ensure long-term sustainability of the site's contents beyond the duration of the Action.

Deliverables/Milestones: Action Website (Yr1)

Task 4. Discipline-specific use cases

Targeted training will be provided for specific research communities through Training Schools. STSMs will be used to enable ESRs to gain experience from and provide feedback to vibrant informatics teams in Europe. Potential synergies will be pursued with existing european training and mobility programmes (e.g. SYNTHESYS, Leonardo, DEST).

Deliverables/Milestones: 4 STSMs (Yr1, Yr2, Yr3, Yr4); 4 TSs (Yr1, Yr2, Yr3, Yr4); WS/H (Yr3)

Task 5. Translating into policy making

In recent years significant progress has been made to identify, through several European funded projects (e.g. BiodiversityKnowledge-KNEU, pro-iBiosphere), the critical elements that facilitate a reciprocal and efficient communication between research and policy making in Europe. These efforts need to be intensified, promote interdisciplinarity and highlight the added value of bringing together information and biodiversity research. Existing reports for policy and decision makers will be updated and collated. This will be done in the context of addressing the technological, socio-cultural, financial and legal impediments for successfully meeting the European societal challenges for 2020.

Deliverables/Milestones: WS (Yr4); OA-P (Yr4); CR (Yr4)

Objective 4: Roadmapping towards a joint research agenda

Task 1. Best practices for e-infrastructure development

Best practice documents and an implementation roadmap are important to introduce clear and precise steps that can be followed. Best practice documents have already been drafted, in the framework of recent European projects and address specific audiences. Within this task existing best practice documents will be gathered and assessed in the context of the Action's challenges and where needed, amendments will be made. This action will draw these documents together to produce an implementation (capacity building) roadmap for key biodiversity e-infrastructures, including service networks, virtual research environments, data modelling tools and data publishing services. The associated deliverables of the Objectives 1,2 and 3 will be used as input for the work in this Task.

Deliverables/Milestones: WS/H (Yr2); OA-P (Yr2); 2 CR (Yr1, Yr2)

Task 2. Domain-specific best practices for data preservation, aggregation and management

The members of the Action have considerable experience in delivering best-practices documents for specific audiences. They also have experience in software documentation. In this Task they will coordinate their activities to consolidate existing best-practices documents and deliver more comprehensive set of documents to be disseminated across the related biodiversity and informatics communities. Special attention will be paid in laying best practice policies for e-infrastructures in relation to data preservation, aggregation and management with regard to infrastructure sustainability.

Deliverables/Milestones: WS/H (Yr4); OA-P (Yr4); 2 CR (Yr3, Yr4)

Summary of outputs

The outputs of this Action are structured to be delivered in an incremental and cumulative way. All the deliverables described are associated with certain tasks within the objectives of the Action, with the exception of the Action conferences that span across the Objectives.

In total the outputs of the Action, based on the initial network of supporters, are summarised below:

- 30 Short Term Scientific Missions (STSM) - At least 50% will be allocated to ESRs. People participating in STSMs are expected to contribute to scientific publications;
- 8 Training Schools (TS) - Training Schools will be organised in regular intervals (two per annum);
- 6 Joined Student Supervisions (JSS) - Joined student supervision of Masters or PhD students;
- 13 Workshops/Hackathons - The exact form will be decided by those assigned to the Task Working Group. Workshops/Hackathons will be organised as side-events of the Action conferences for more efficient use of available resources and for maximising on-site interactions;
- 28 Open Access Publications (OA-P) - Scientific publications will be compiled by the broadest possible author base;
- 10 Consolidated Reports (CR) or Task-specific documents;
- 2 Conferences - Meetings will be co-organised with other European or nationally funded projects or organisations for financial efficiency and for increasing the number of interactions;
- The Action's website

Furthermore, the network of the Action will identify potential funding opportunities within the Horizon 2020 Programme or through national and regional funding calls, in order to secure the expansion of the network and the continuation of its activities.

Action structure and participation – Working Groups, management, internal procedures

Management Committee (MC)

After its formulation, the MC will convene for the first time during the kick-off meeting of the Action, to decide on the synthesis of the rest of the management and operational instruments, including the composition of a Steering Committee (StC), Working Groups (WG) and Action Activity Officers (AAO). The MC will be responsible for assessing the overall progress of the Action and make necessary adjustments to increase the efficiency of internal procedures. The MC will physically meet at least once per year. When possible, MC meetings will be carried out during Action conferences.

Working Groups (WG)

The WGs will be created against the challenges described under Section A. They will act as the primary vehicle for driving forward the Action objectives and meeting its deliverables and milestones. The composition of all WGs will be decided in the context of achieving the maximum possible diversity, in terms of disciplines, experience, geographic coverage and gender. For each of the WGs a scientific coordinator and a deputy-coordinator will be appointed. A rapporteur will also be appointed with the responsibility of summarising and codifying meeting proceedings. Rapporteurs from each of the WGs will also be encouraged to participate in the meetings of the rest of the WGs (by either physical or virtual presence) acting as liaisons between the WGs.

WG1: *Assessment of existing models and standards for biodiversity data sharing*

This WG will primarily focus on achieving the deliverables described under the Objective 1 of the Action. The WG will have a highly cross-domain and interdisciplinary composition to address the complexity of assessing existing models and standards in the scope of the needs of taxonomic, genomic, and ecosystem research.

WG2: *Data availability*, *interoperability and harmonisation*

This WG will focus on achieving the deliverables described in Objective 2 of the Action. The Group will work closely together with WG1 and will use WG1 proceedings as primary input. The WG members will act as the primary members of the organising committees of the Action conferences.

WG3: *Education*, *training and dissemination*

This Group will align its activities with Objective 3. The group will decide on the strategic direction of STSMs, Joint student supervision, Training Schools and workshops. The Group will develop communication strategy for communicating outputs of the Action to decision and policy makers. Furthermore, it will initiate the development and sustain the Action's

website. The Group will work closely together with both AAOs (see related section below) of the Action.

WG4: *Data integration for cross-disciplinary research*

This WG will work focusing on the deliverables of Objective 4. The Group will approach biodiversity informatics in an inclusive and comprehensive way. Members of the WG will interact with the rest of the WG rapporteurs and coordinators to incorporate the output of their respective Groups. The Group will develop detailed and practical guidelines for each of the major biodiversity related disciplines and stakeholders

Action Activity Officers (AAO)

Action Activity Officers will organise and monitor the progress of the specific activities related to the objectives of the Action. AAOs will participate, ex-officio, in the Action's StG, will monitor the progress of related activities, deliver reports of work and propose course of actions according to the Action's objectives.

a. AAO for Education & Training

The Education & Training AAO will contribute to the organisation of the STSMs, Training Schools, Workshops and Joined Student Supervisions. The AAO will act as a contact point for trainees and liaise between institutions to facilitate researchers' mobility between different research groups.

b. AAO for Dissemination & Outreach

The Dissemination & Outreach AAO will liaise between all Action's participants to monitor and generate metrics on dissemination and outreach activities, overlook the Action's website initial development and updating, and prepare documents for the general public and the press.

Participation model

The participation model for the initial network is balanced, ensuring: (i) a wide geographic spread across European countries that will support the Action's vision to create the first pan-European network on biodiversity informatics and (ii) the wide participation of partners from institutions with substantial experience in either of the scientific domains and disciplines this Action crosses, including universities, research institutions, natural history museums, international organisations, NGOs and SMEs. This safeguards the ability of the network to deliver in time and in scope.

The initial network is comprised of 65 proposers from 24 COST Countries, two Near Neighbor Countries (NNC) and one International Partner Country (IPC). The spectrum of expertise behind the people involved spans from environmental and marine biology, ecology, plant biology, botany, zoology, comparative biology and natural history to bioinformatics, databases, data mining, data curation, computational modelling, theory of

scientific computing and data processing. This diverse group of people ensures a dynamic head start for the Action and highlights the trans-domain nature of BioUnify. After the approval of the Action, further work will be done in attracting participants from across Europe, with emphasis on dynamic groups from Countries with limited access to funding for networking and collaboration activities.

During the Action, provisions will be made for gender, experience, expertise and geographic coverage balance across all training and dissemination activities, as well as the participation in the Action's management structures.

Funding program

COST Open Call - ID: oc-2014-1

References

- Hardisty A, Roberts D, Community TBI (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13 (1): 16. DOI: 10.1186/1472-6785-13-16
- Hobern D, Apostolico A, Arnaud E, Bello JC, Canhos D, Dubois G, Field D, García EA, Hardisty A, Harrison J, Heidorn B, Krishtalka L, Mata E, Page R, Parr C, Price J, Willoughby S (2013) Global Biodiversity Informatics Outlook: Delivering biodiversity knowledge in the information age. Global Biodiversity Information Facility (Secretariat), Copenhagen, 44 pp. [In English]. URL: http://www.biodiversityinformatics.org [ISBN 87-92020-52-6]
- P. Bryan Heidorn (2008) Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57 (2): 280-299. DOI: 10.1353/lib.0.0036