 doi:10.20944/preprints202009.0357.v1

A Streamlined Workflow for Conversion, Peer-Review and Publication of Omics Metadata as Omics Data Papers

Mariya Dimitrova[1*], Raïssa Meyer[2], Pier Luigi Buttigieg[3], Teodor Georgiev[4], Georgi Zhelezov[5], Seyhan Demirov[6], Vincent Smith[7], Lyubomir Penev[8]

[1] Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria;
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev St., Block 25A, 1113 Sofia, Bulgaria
Correspondence address: Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Email: m.dimitrova@pensoft.net
https://orcid.org/0000-0002-8083-6048
*Corresponding author

[2] Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany
Correspondence address: Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany
Email: raissa.meyer@awi.de
https://orcid.org/0000-0002-2996-719X

[3] Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany
Correspondence address: Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Bremerhaven, Germany
Email: pier.buttigieg@awi.de
https://orcid.org/0000-0002-4366-3088

[4] Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Correspondence address: Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Email: t.georgiev@pensoft.net
https://orcid.org/0000-0001-8558-6845

[5] Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Correspondence address: Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Email: g.zhelezov@pensoft.net

[6] Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Correspondence address: Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Email: programmer@pensoft.net

[7] The Natural History Museum, London, United Kingdom
Correspondence address: The Natural History Museum, London, United Kingdom
Email: vince@vsmith.info
https://orcid.org/0000-0001-5297-7452

[8] Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria;
Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, 2 Gagarin Street, 1113 Sofia, Bulgaria

 doi:10.20944/preprints202009.0357.v1

Correspondence address: Pensoft Publishers, Prof. Georgi Zlatarski Street 12, 1700 Sofia, Bulgaria
Email: l.penev@pensoft.net
https://orcid.org/0000-0002-2186-5033

 doi:10.20944/preprints202009.0357.v1

Abstract

Background

Data papers have emerged as a powerful instrument for open data publishing, obtaining credit, and establishing priority for datasets generated in scientific experiments. Academic publishing improves data and metadata quality through peer-review and increases the impact of datasets by enhancing their visibility, accessibility, and re-usability.

Objective

We aimed to establish a new type of article structure and template for omics studies: the omics data paper. To improve data interoperability and further incentivise researchers to publish high-quality data sets, we created a workflow for streamlined import of omics metadata directly into a data paper manuscript.

Methods

An omics data paper template was designed by defining key article sections which encourage the description of omics datasets and methodologies. The workflow was based on REpresentational State Transfer services and Xpath to extract information from the European Nucleotide Archive, ArrayExpress and BioSamples databases, which follow community-agreed standards.

Findings

The workflow for automatic import of standard-compliant metadata into an omics data paper manuscript facilitates the authoring process. It demonstrates the importance and potential of creating machine-readable and standard-compliant metadata.

Conclusion

 doi:10.20944/preprints202009.0357.v1

The omics data paper structure and workflow to import omics metadata improves the data publishing landscape by providing a novel mechanism for creating high-quality, enhanced metadata records, peer reviewing and publishing of these. It constitutes a powerful addition for distribution, visibility, reproducibility and re-usability of scientific data. We hope that streamlined metadata re-use for scholarly publishing encourages authors to improve the quality of their metadata to achieve a truly FAIR data world.

Keywords

Data； data paper; omics; metadata; workflow; standards; FAIR principles, MIxS; MINSEQE

1. Introduction

The term “omics” refers to the study of biological systems through the examination of different elements of the molecular basis of life. For example, the genome is examined through the analysis of gene (DNA) sequences, the transcriptome is the collection of all mRNA molecules in an organism, and the metabolome is the collection of all metabolites and intermediate substrates participating in the metabolic pathways. Omic studies are generating large quantities of deeply minable data with increasing scale and complexity [1, 2]. Further, omics technologies and approaches have revolutionised biodiversity science [3, 4, 5].

Independently from the recent development of omics technologies and data generation, however, the publication of high-quality omics biodiversity data and its accompanying, standardised metadata, is still neither harmonised nor interoperable

 doi:10.20944/preprints202009.0357.v1

[6]. Existing infrastructures in omics data science focus on the sequence or molecular data generated from omics studies. The databases of the International Nucleotide Sequence Database Collaboration (INSDC) [7, 8] have provided a trusted archive for these data. In parallel, major infrastructures to handle higher-order biodiversity data (e.g. occurrences linked to taxa, specimen records) have emerged and include the Global Biodiversity Information Facility (GBIF) [9], the Integrated Digitized Biocollections (iDigBio) [10], the Distributed System of Scientific Collections (DiSSCo) [11], the Ocean Biogeographic Information System (OBIS) [12], the Global Genome Biodiversity Network (GGBN) [13] and others. Some of these infrastructures support data repositories which follow community-accepted metadata standards. GBIF uses the Darwin Core Standard (DwC) for biodiversity data, while the GGBN have developed their own GGBN Data Standard, which interoperates with DwC and the Access to Biological Collections Data (ABCD) schema for primary biodiversity data [14, 15, 16, 17]. The INSDC cooperates with community standards initiatives such as the Genomics Standards Consortium (GSC) to implement their Minimum Information about any (x) Sequence (MIxS) standard for genomic, metagenomic and environmental metadata descriptors and with the Global Microbial Identifier (GMI) group for pathogen sequence metadata [18, 19, 20]. MIxS consists of three checklists each containing several packages for the description of various environments where genomic material could be sampled from. Other international data repositories such as EBI EMBL's ArrayExpress and the BioSamples database implement standards such as Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE) and Minimum Information About a Microarray Experiment (MIAME) and various MIxS environmental checklists [21, 22,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

23].

The presence of a digital infrastructure and standards supporting a certain data class is just one of the necessary conditions for adequate data sharing and re-use, which is often impeded by the insufficient use of these standards and inadequate incentives for the stakeholders participating in the process [8]. The concept of Findable, Accessible, Interoperable and Re-usable (FAIR) data is a major step forward in building the foundation for sharing reusable data [24]. How can we, however, incentivise the data creators, holders, scientists and institutions to pursue the FAIRsharing of data, information, and knowledge exchange [25] in omic biodiversity science?

There are different ways scientists can publish their data, however, all can be attributed to two main routes: (1) data publishing through international trusted data repositories, such as INSDC [7], GBIF [9], and others, and (2) scholarly data publishing in the form of data papers or as data underpinning a research article [26, 27, 28, 29, 30]. While the first route focuses on data aggregation, standardisation and re-use, the second one augments the quality and reusability of data and metadata through peer reviewing and data auditing in the scholarly publishing process. Scholarly data publishing provides an opportunity to enhance the original metadata in the data paper narrative and to link it to the original dataset via stable identifiers, thus improving the reproducibility and findability of the data. Furthermore, it ensures a scientific record, crediting and acknowledgement for the data creators and scientists in the form of citable scholarly articles. Academic publishing involves

 doi:10.20944/preprints202009.0357.v1

dissemination of research through additional channels, such as journal distribution networks, and creates increased opportunities for open science collaboration [26].

While standards and infrastructures are crucial to the advancement of data sharing and reuse within the field of omics, we argue that incentivising authors to publish their data in the form of peer reviewed journal articles (data papers) creates the driving force towards a truly FAIR data world.

As more and more researchers want to deposit and share their datasets, standards, infrastructures, and workflows become central to delivering FAIR data. Following the example set by Chavan and Penev [26], who introduced data papers in biodiversity science, we have established a concept for an omics data paper - a type of scholarly paper in which data, generated in genomic or other omic experiments, is described with extended and peer reviewed metadata, and linked to the corresponding dataset(s) deposited in an INSDC or other archive. To further incentivise authors to publish omics data papers and to demonstrate the importance of high-quality metadata, we propose a streamlined workflow for conversion of European Nucleotide Archive (ENA) metadata directly into a data paper manuscript.

The aim of the present paper is to conceptualise the omics data paper, and to describe the streamlined workflow for automated import of metadata into an omics data paper manuscript. This workflow also accommodates the peer review and publication processes associated with the manuscript.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

2. Methods

Approach

We took the following steps to approach the goal of establishing an omics data paper template and workflow:

1. Identify the high-level needs of the omics communities.
2. Review existing standards, infrastructure [32] and datasets, as well as the existing data paper formats for describing omics datasets [33, 34, 35].
3. Synthesise the technical solutions and incorporate further functional needs to create the structure of the new type of data paper.

We created a template, defining article sections and subsections to map the article narrative to metadata associated with the dataset(s) described in an omics data paper.

Workflow for extracting relevant metadata from ENA XML files

We proceeded to develop a workflow for automatic import of metadata into omics data paper manuscripts based on ENA's metadata structure, as well as the ArrayExpress [21] and BioSamples [22] databases. The workflow uses REST API requests and Xpath to retrieve segments of information from XML files from ENA, ArrayExpress and BioSamples [36]. It then imports them into our proposed data paper manuscript structure.

For demonstration, testing and reproducibility purposes, this workflow was implemented in a R shiny app [37, 38, 39] which visualises metadata extracted from

 doi:10.20944/preprints202009.0357.v1

ENA inside the relevant sections of the proposed manuscript template within the application interface. During this prototyping stage, we tested the app with multiple ENA study identifiers to guide the improvement of the import algorithm. The R shiny app is available both as open source code on Github [40] and as an interactive web app [41] deployed in an RStudio cloud environment [42] and hosted on a Shinyapps.io server [37, 38]. The R version at the time of developing the R Shiny app was R version 4.0.0 (2020-04-24) (Arbor Day) [39].

Integration of metadata extraction workflow with the ARPHA Writing Tool

The metadata extraction workflow was completed with a conversion tool working "inside" the Pensoft's ARPHA Writing Tool [31]. A new type of publication and template, "OMICS data paper", was created, following the proposed data paper structure. Certain sections of the omics data paper template were made mandatory such as the "Methods" section and the "Data resources" section.

An important component of the design and implementation of the omics data papers is the BioSamples Supplementary Table. ENA metadata records that contain links to associated BioSamples metadata (MIxS checklists) [20, 22] are retrieved by the automatic import workflow, and are transformed into a long format table, which will be attached to the manuscript as a comma-separated value (CSV) file. We restrict editing of Supplementary Tables imported from BioSamples to prevent metadata loss and tampering. Authors can only change the MIxS checklists related to their manuscript if they re-upload them to BioSamples. Synchronisation with BioSamples

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

from the manuscript in the ARPHA Writing Tool is enabled through a button labelled "Re-import from BioSamples".

3. Findings

Structure of the OMICS data paper

The omics data paper describes datasets generated in omics research. The described dataset is at the core of the data paper, but the methodology required to obtain it is just as valuable as the data itself. To guide authors in the authoring process and to better inform the readers about the contents of the proposed data paper, we designed a detailed manuscript template (Table 1).

Article section	**Purpose**	**ENA metadata source field**
Abstract	Summary of the value of the study, the experimental design and the dataset itself.	**Study/Project XML:** //abstract

Introduction
- Value of the dataset
  - Scientific value
  - Societal value
	Outline of the reason for the study. Authors should put into perspective its value for the scientific and broader communities. Often sequencing studies are part of large scale genome sequencing projects and this article section allows authors to explain their role in them.	*Written by the authors*

Methods
- Sampling
  - Environmental profile
  - Geographic
	This section is split into 3 major parts to describe how the physical material was collected, processed and transformed into a	**ArrayExpress XML> Protocol XMLs:** protocol/type protocol/text protocol/hardware

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

range
- Technologies used
- Sample processing
- Technologies used
- Data processing
	dataset.
	The “Sampling” section allows authors to outline the environmental and geographic characteristics of the locations where their material was collected. Authors are encouraged to share as much detail as they can (e.g. geographic coordinates, habitats, seasonal information, etc.). The sampling methods should be described in the “Technologies used” subsection.
	“Sample processing” should explain the laboratory procedures involved in the transition of the physical sample into its digital footprint. Finally, the “Data Processing” subsection should mention the steps taken to transform the raw dataset into the one which was published (e.g. normalisation steps).
	None of the subsections are compulsory and the authors can write the Methods in a form outside these topics but our template provides a detailed best practices structure to follow.
		protocol/software
		
		And
		
		Experiment XMLs:
		//EXPERIMENT/DESIGN/LIBRARY_DESCRIPTOR/LIBRARY_STRATEGY
		
		And
		
		Experiment XMLs:
		//EXPERIMENT/PLATFORM (-> Sample processing/Technologies used)
		
		And
		
		Sample XMLs:
		
		//SAMPLE/DESCRIPTION
		//SAMPLE/SAMPLE_ATTRIBUTES/SAMPLE_ATTRIBUTE

Biodiversity profile - Target - Taxonomic range - Functional range - Traits	This section describes the experimental design of the study. The target refers to the molecular target being studied (i.e. DNA, RNA, protein). The taxonomic range refers to the taxonomy of the studied organism(s) or the	*Written by the authors*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

	taxonomic composition of a metagenomic sample. The authors are encouraged to use a common taxonomy but they can also provide their own during the authoring process in AWT. Authors can specify a particular range of biological functions which was the subject of their study (e.g. metabolic functions), as well as specific traits (e.g. pathogenicity) if relevant to the study.	
Data resources	This is the section which contains a link to the dataset(s) (preferably to its permanent resolvable identifier, such as a DOI), as well as any accession numbers and data formats.	**Study/Project XML:** //XREF_LINK/ID[../DB='ENA-FASTQ-FILES']
Data statistics	Quantitative and qualitative description of the dataset. (e.g. read depth, coverage, base ratios). This section helps readers to quickly evaluate the dataset by gauging some of its characteristics without having analysed the dataset themselves. Some of the data statistics can be represented as charts and/or short tables.	*Written by the authors*
Caveats and limitations	A section to discuss what could be improved in the experiment, what future steps could be taken and what to consider when re-using the published data.	*Written by the authors*
Usage rights	Rights and licenses to use the data. The data paper	*Written by the authors*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

	is open access by default.	
Supplementary table	Contains imported MIxS checklists for the imported BioSamples. The checklists are in long format. The table can be downloaded as a separate comma-separated value (CSV) file after publication.	**Sample XMLs:** //SAMPLE/IDENTIFIERS/EXTERNAL_ID[@namespace="BioSample"] And **BioSample XMLs:** //Property[@class]

Table 1. OMICS data paper sections, their purpose and ENA metadata fields from which they are populated, if such fields exist. Values in the third column refer to the fields in ENA's XML files which contain the information used to automatically fill in the relevant section of the template. We have pointed to the type of XML (marked in bold) as well as the Xpath used to extract the information. Some data paper sections do not have ENA metadata fields associated with them and the authors are encouraged to fill in their contents.

The template focuses on the value of the data, the methods used to generate it and the qualitative and quantitative characteristics of the dataset.

Omics metadata extraction workflow

Metadata describing the datasets was utilised to facilitate creation and authoring of the data paper manuscript. By following ENA's metadata model [32], including its links to the ArrayExpress [21] and BioSamples [22] databases, we designed a workflow which orchestrates the extraction of relevant metadata from the various ENA XML files (Fig. 1). The Study XML and the Project XML are the starting points

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

in the proposed workflow as they integrate all other types of data and metadata available in ENA for a given scientific study. Each metadata object in the ENA metadata model is associated with a unique identifier, which can be used to retrieve its corresponding XML file via the ENA API [32].

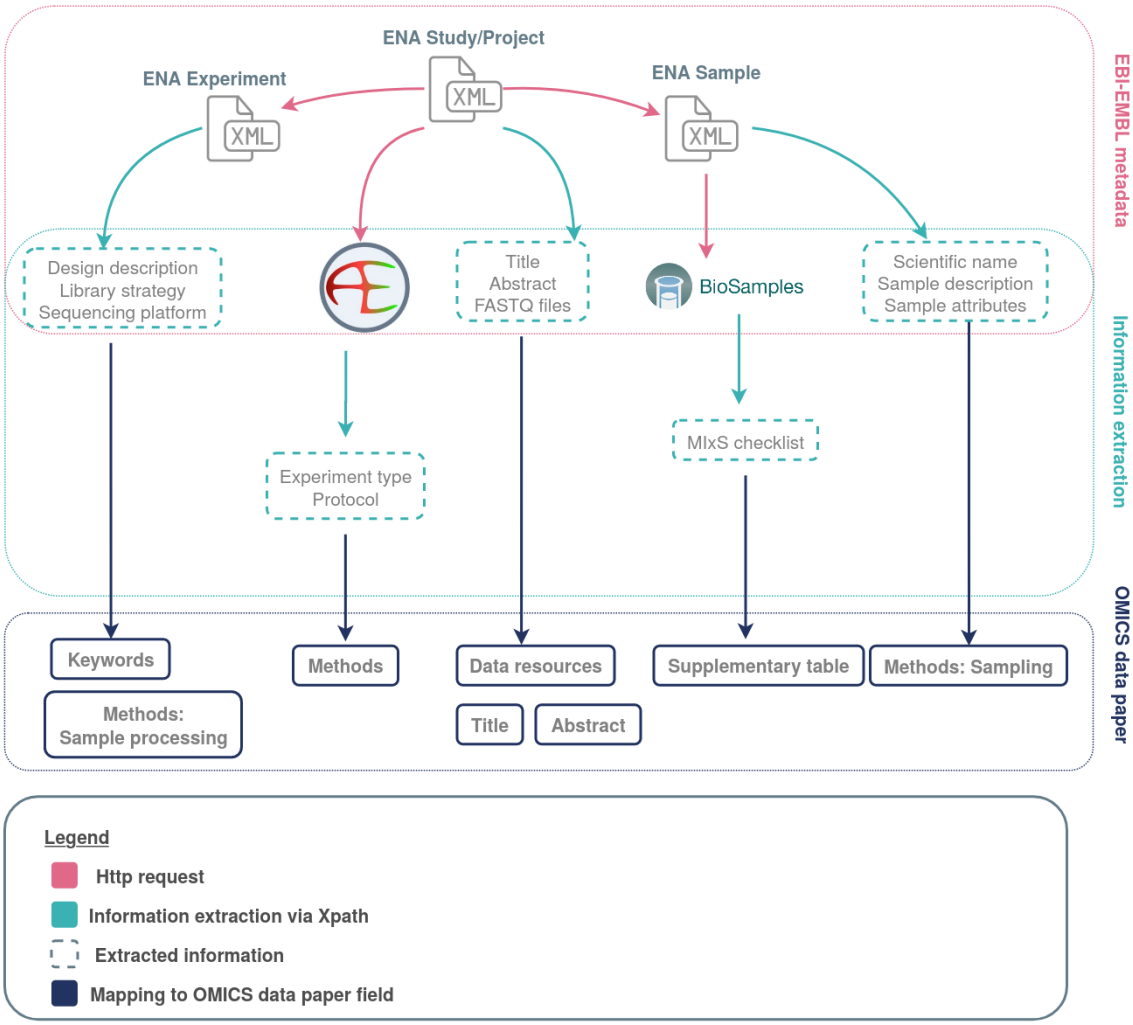

Fig. 1 Metadata extraction workflow from ENA, ArrayExpress and BioSamples

As outlined in our proposed workflow (Fig. 1), the Study or Project accession number is used to obtain a XML file which contains the accession numbers for all associated

 doi:10.20944/preprints202009.0357.v1

Experiment and Sample metadata objects, and in some cases ArrayExpress and BioSamples metadata objects.

ArrayExpress is a database storing data and metadata from functional genomic microarray or sequencing experiments [21]. ArrayExpress' own submission platform Annotare and curators ensure that metadata from all sequencing experiments follow the Minimum Information About a Sequencing Experiment (MINSEQE) standard [21, 23].

Raw data from sequencing experiments submitted to ArrayExpress are also automatically deposited in ENA [43] as part of a Study metadata object, linked to Experiment and Sample objects [32]. Provenance of metadata imported from ArrayExpress can be established through a unique ArrayExpress accession number in the ENA Study XML. We integrated the extraction of curated, MINSEQE compliant metadata from ArrayExpress into the workflow, thus enhancing manuscripts with high quality metadata about experimental design and methodologies.

Another database within EMBL-EBI's infrastructure is BioSamples, a database which "stores and supplies descriptions and metadata about biological samples" [22]. Metadata descriptors in BioSamples records follow the MIxS standard [22]. Depending on the type of sample, submission to BioSamples requires different MIxS checklists to be filled in, after which they are publicly available in the form of XML files [22]. Unique identifiers link Sample XMLs from ENA with their associated BioSamples XML records. Thus, we are able to extract BioSamples information for any samples from a given ENA Study or Project. BioSamples records are imported

 doi:10.20944/preprints202009.0357.v1

into a table which is attached to the manuscript as a Supplementary CSV file. This supplementary table is a mandatory component of a manuscript, when accompanying BioSamples records are available, and cannot be removed by the authors. In cases when the authors spot a mistake in their submitted metadata, they are encouraged to change it within the BioSamples database. Upload of standalone BioSamples MiXs checklists is not permitted in the ARPHA Writing Tool, so that authors perform their corrections in the original metadata repository. After that, they can automatically retrieve them from BioSamples and import them into the manuscript with the click of a button. Thus, we promote the reuse and interoperability of MIxS compliant metadata sourced from BioSamples.

We implemented the template and workflow into Pensoft's ARPHA Writing Tool [31] , enabling import of the extracted ENA metadata records into the omics data paper template (Table 1). Fig. 2 shows a diagram demonstrating the import functionality from the perspective of the user.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

Fig. 2. Automatic metadata import from ENA, ArrayExpress and BioSamples, facilitates the creation of omics data paper manuscripts inside the ARPHA Writing Tool.

 doi:10.20944/preprints202009.0357.v1

R shiny app - deployment and reproducibility

The template and workflow were first prototyped in a R shiny app [41], the code for which is open source and available on Github under Apache 2.0 license [40], as outlined in the Methodology section of this paper. The R shiny app is a web application emulating the functionality of the metadata import workflow in the ARPHA Writing Tool. The application runs in a virtual R environment [39, 42] and is deployed and hosted on the web via Shinyapps.io [37, 38], configured to allow up to 50 concurrent connections.

The interface of the application features a text field for input of ENA Study or Project ID and an ‘Convert’ button controlling the import of metadata and conversion to manuscript. Three buttons to download the outputs appear after the ‘Convert’ button is clicked. The generated manuscript narrative, along with a data frame containing the BioSamples MIxS checklist, are visualised in the R shiny app interface. The narrative can be downloaded as a HTML file by clicking the ‘Download HTML’ button, whereas the BioSamples checklist can be downloaded as a CSV file by clicking the ‘Download Supplementary Material’ button. This CSV file is identical to the one generated as a supplementary file by the ARPHA Writing Tool.

The R shiny app has one additional functionality, which is not present in the workflow implemented in the ARPHA Writing Tool: it transforms the imported metadata into a Journal Article Tag Suite (JATS) XML file [44], which can be downloaded by clicking the ‘Download XML’ button. We validated the XML against the latest JATS DTD version with the JATS4R validator [45]. The JATS XML is structured according to the

 doi:10.20944/preprints202009.0357.v1

Pensoft omics data paper template so that most article section nodes are defined with the sec tag and an attribute sec-type is used to define the exact section name (e.g. the Methods section is marked in the XML as <sec sec-type="Methods">). A basic “skeleton” file of the JATS XML file is available in the project Github repository [40].

Despite being tailored to the Pensoft omics data paper template, JATS XML files generated via the R shiny app can be used by other publishers or individuals to generate their own omics data paper manuscripts. Together with ENA’s documentation about programmatic access to its resources [36], the codebase enables reproducibility of our workflow and creates the potential for it to be deployed by other journals or publishers.

4. Discussion

The data and metadata publishing landscape

The concept of data papers is not new; in fact, they have been in existence for several decades already. One of the first journals to implement this concept was Ecological Society of America’s Ecological Archives [46, 47]. In 2011, Chavan and Penev envisioned metadata as a resource for authoring data papers for primary biodiversity data and identified a lack of clear guidelines and good practices for authoring metadata (the “how”) and the incentives for authors to do so (the “why”) [26]. They proposed data papers as a “mechanism to incentivise data publishing in biodiversity science” and introduced them to the biodiversity community through Pensoft’s journals. To further simplify data paper authoring, Pensoft pioneered an integrated workflow for automatic metadata-to-manuscript conversion of primary

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

biodiversity datasets published through GBIF's Integrated Publishing Toolkit (IPT) [26, 29, 48].

This streamlined metadata conversion workflow was first introduced in several of Pensoft's biodiversity journals and then in other journals by other publishers, such as Nature's Scientific Data, PLOS ONE, BMC Ecology and many others [49]. Since 2011, nearly 300 data papers have been published in Pensoft's journals and the average number of published data papers continues to grow (Fig. 3). The total number of data papers published by other publishers is in the thousands (Fig. 4) [50]. Data papers are no longer an abstract idea but have already been practically implemented in multiple journals in different disciplines.

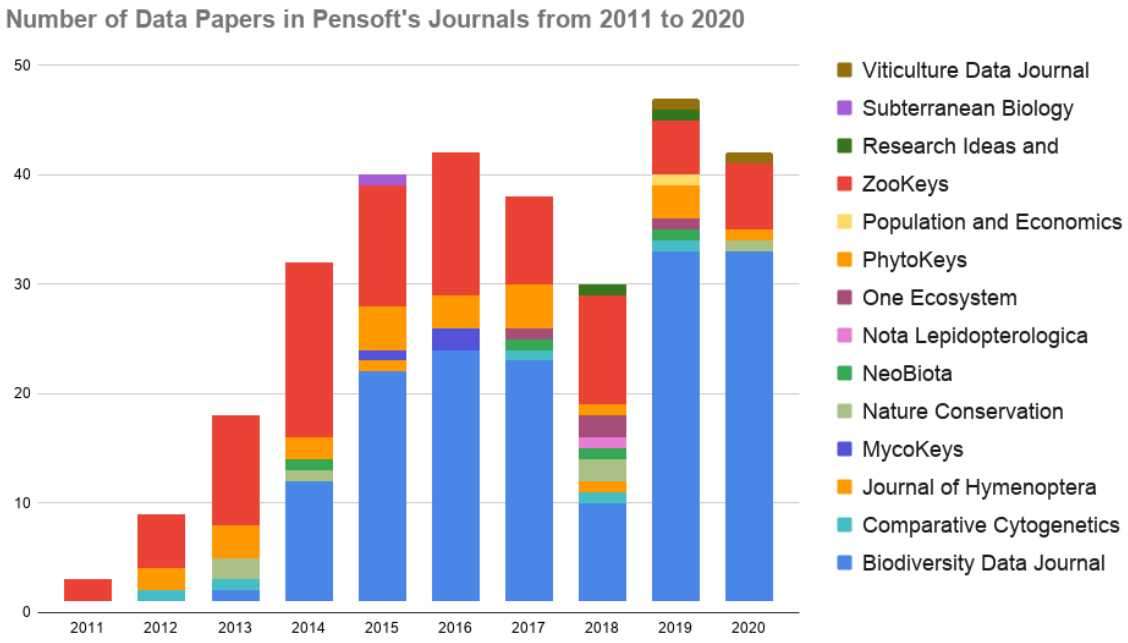

Fig 3. Numbers of published data papers in Pensoft journals from 2011 to 2020.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

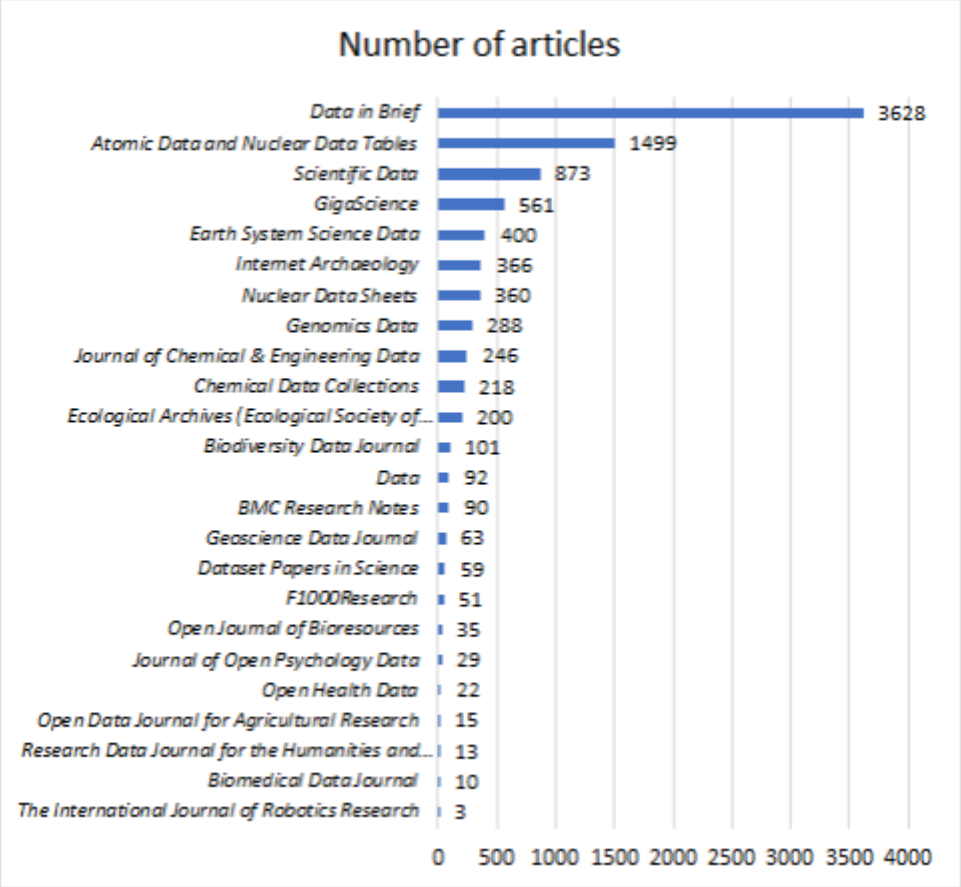

Fig. 4 Number of data papers published by different journals in the fields of science and the humanities [50]. Permission to use the figure was obtained from Dr. Joachim Schöpfel, leading author of [50].

Comparison with other tools and workflows

Since 2011, Pensoft has developed other integrative ways to streamline metadata authoring and data paper publication by integrating different workflows into its collaborative online authoring tool, the ARPHA Writing Tool (AWT) and associated Biodiversity Data Journal [51]. For instance, Ecological Metadata Language (EML) metadata files used in the IPT can be directly converted and imported into manuscripts in AWT “at the click of a button”, then edited in the tool and submitted to the Biodiversity Data Journal [30, 52]. This workflow closely resembles the workflow described in this paper but it is focused on ecological data. The EML workflow

 doi:10.20944/preprints202009.0357.v1

accepts a single specimen record identifier and imports information about that record from several infrastructures (GBIF, Barcode of Life Data Systems (BOLD), iDigBio, or PlutoF) into manuscripts [52]. It also enables conversion of an EML-formatted file into a biodiversity data paper [52], a functionality not covered by the present workflow, which only performs API requests.

Generation of extended metadata descriptors has been the focus of other tools, such as the Metadata Shiny Automated Resources and Knowledge (MetaShARK) [53] and Datascriptor [54], which is still under development. MetaShARK aims to facilitate assembly of ecology metadata by providing a user-friendly workflow for metadata packaging [53]. Unlike the workflow described here, it is more focused on primary metadata generation than metadata sharing and reuse [53]. Our workflow uses already generated metadata and provides a template for their extension to create an extended metadata description converted to narrative. Datascriptor is more closely related to our workflow because it aims to transform metadata, generated by following community standards, into a data article [54]. To do so, the developers have envisioned the generation of a JATS XML [54], which is what we have implemented in our R shiny app demonstrating the workflow for import of metadata into omics data paper manuscript.

Data papers for the field of omics: rationale and benefits

Generation of omic data and metadata is one of the very first outputs of the research cycle, but not all of this is shared via research publications. Even when these data are published, the focus is usually on the interpretation of the data, rather than

 doi:10.20944/preprints202009.0357.v1

metadata quality or the FAIR properties of the dataset. Deposition of raw omic data, such as sequencing data, mass spectrometry (MS) proteomic data and RNA-sequencing data, into centralised databases has become a routine practice for studies involving omic experiments [55]. ENA provides the necessary infrastructure to share sequencing data in a structured format and enables machine-readability and interoperability through the use of identifiers, consistent schema models and APIs [32]. However, for the data to be communicated to and shared with other researchers it needs to be described inside a human-readable narrative. With our proposed omics data paper and the automatic import workflow, we encapsulate all metadata about a study into a single piece of narrative, thus completing the scientific process.

Authoring omics data papers, despite being aided by the automated workflow, requires additional effort and time. Here we outline some of the benefits which make the process of creating such manuscripts worthwhile, as well as on how data interoperability contributes to the FAIR data and metadata publishing landscape.

1. *Omics data papers and underlying datasets undergo peer-review and data auditing*

   Prior to peer-review of submitted omics data paper manuscripts, all underlying datasets go through mandatory data auditing, cleaning and quality checks to assure that they meet the journal standards for publication [56]. This is done by a data auditor, whose role is to technically evaluate the submitted datasets for compliance to a data quality checklist [57] and to provide authors with a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2020 doi:10.20944/preprints202009.0357.v1

detailed report, including recommendations for improving the dataset. Only after the authors change the dataset according to the recommendations, it can be approved for peer-review. The introduction of data scientists into the publishing process ensures that submitted data and metadata are FAIR, consistent and of high-quality. This double checking - first of the datasets by the data auditors and then of the whole manuscript by the reviewers - is an meticulous approach to enhancing the quality of datasets and to the best of our knowledge has not been adopted by any other publisher so far.

2. *Publication of data papers improves metadata quality*

Authoring metadata is a necessary step to publish omics data into an open repository; however, there is considerable variability when it comes to the quality of the published metadata [6, 55]. The workflow allows metadata authors, metadata standard creators and data repository managers to evaluate the quality of metadata files deposited to INSDC databases. Throughout our testing phase we came across many datasets with missing or incorrectly formatted metadata fields. A recent observation by members of the Genomics Standards Consortium found that missing or incomplete metadata records from SARS-CoV-2 genomic and metagenomic studies are of frequent occurrence in INSDC databases and other repositories: such deficits of high-quality, community-standard metadata, have become apparent during the COVID-19 healthcare crisis as global scientific efforts have been directed at generating and analysing data related to the novel coronavirus and data sharing and re-use have become crucial [6].

 doi:10.20944/preprints202009.0357.v1

Curation of metadata, currently implemented by ArrayExpress via their Annotaire tool [58, 59], is an adequate method for high-quality metadata publishing, based on standards. Most important, however, is that metadata authors learn to adopt and correctly use the existing standards in the process of describing their data. By directly observing the role of their metadata in creating the manuscript, they are made aware of its value and should be incentivised to improve the quality and quantity of the metadata they provide. After importing metadata into their omics data paper manuscript, authors would need to manually correct and fill in the missing information, which defies the main purpose of the workflow: to make the data better described through extended and detailed metadata in the form of peer-reviewed, widely accessible and citable data papers.

3. *High-quality metadata enables data-driven discovery*

Metadata which follows community accepted standards is vital for data-driven discoveries as it provides the necessary context to characterise the dataset it describes. Omics data papers do not only improve the quality of the metadata but also constitute an enhanced metadata record themselves. As a result, the publication can inspire new research ideas and open new possibilities for use of the dataset. For instance, in pharmaceutical science, old compounds are commonly researched as part of the development of new drugs because they could harbour unexplored biological activities [60]. By giving further visibility to omic datasets through their publication in an omics data paper and by

enhancing metadata through publication, we stimulate scientific research and data-driven discovery.

4. *Data papers help to establish priority*

Publishing data papers at early stages of the research process can provide an important benefit for authors: the opportunity to get the first scientific record for their effort in assembling a dataset and obtain feedback from the research community. It is well known that many authors are hesitant to publish datasets which they have not yet analysed or used for supporting any research findings for fear of someone else using the data and getting 'scooped'. By publishing a data paper, the authors are guaranteed that the described data can be re-used in accordance with the Open Science principles, following all community accepted ethical norms for citation, priority and generating new knowledge through joint publications based on shared data.

5. *Publishing omics data papers is a way to obtain credit for one's work*

Science crediting further incentivises researchers to publish omics data papers because their work impact can be measured in a way familiar to authors of traditional research papers, adding to their researcher impact metrics. In addition, the data managers and scientists who generate the data are not always among the authors of traditional research articles, which focus on the data analysis and outcomes. Thus, data paper publishing can be a way for all actors involved in the process of gathering, curating and managing the

 doi:10.20944/preprints202009.0357.v1

data - be they early stage researchers, technicians or data scientists - to obtain credit for their valuable work.

Limitations and future outlooks

The automated workflow for importing omics metadata into data paper manuscripts currently works only with ENA metadata records. While INSDC metadata is exchanged across all three databases in the consortium (ENA, GenBank and DDBJ) [7], it would be beneficial if users could import metadata from any of the three data repositories via their associated identifier. The reason for the current limitation is the requirement for additional integrations, produced by the variation of APIs and the differing metadata schemas. We decided to integrate the workflow with ENA as the first showcase of this novel method of creation of data paper manuscripts, because their metadata structure was easiest to work with and because they share data with GenBank and DDBJ.

Currently, the streamlined metadata import workflow for the omics data paper is focused mostly on genomic data. In the future, we plan to expand the workflow to include other repositories and data types, such as metagenomics data and operational taxonomic units (OTU) tables. This addition will integrate new data science solutions for efficiently and interoperably exchanging and storing sparse and high dimensional contingency tables along with their associated sample and taxonomic metadata (e.g. the BIOM format [61]). Thus, we support the development away from the fragmentation of data and towards a single quantum of information to exchange, containing interoperable, accessible, and transparent information. Making

 doi:10.20944/preprints202009.0357.v1

use of this advancement, future workflows for omics data paper creation may also support BIOM files for data provision, as outlined in an unpublished dissertation by Raissa Mayer (2020).

Integrations between existing infrastructures and data-driven initiatives are key to the FAIRness of data and metadata. The streamlined workflow for import of metadata from ENA, ArrayExpress and BioSample is another step in this direction. However, to make metadata truly FAIR, there should be a two-way link between the original data and metadata repository (e.g. ENA) and the enhanced metadata record (e.g. the omics data paper).

5. Conclusions

In conclusion, the new omics data paper, implemented in Pensoft's publishing process provides a mechanism for incentivising omics data sharing and reuse through scholarly publishing. In addition, the workflow for import of metadata into manuscripts encourages and incentivises authors to enhance data quality and completeness. The workflow also demonstrates the importance of linking data from different infrastructures using stable identifiers and thus sets an example for future integrations with other metadata and data repositories.

Availability of supporting source code and requirements

- Project name: Omics Data Paper Generator
- Project home page: https://github.com/pensoft/omics-data-paper-shinyapp
- Operating system(s): Platform independent

- Programming language: R
- Other requirements: R version 4.0.0 (2020-04-24) (Arbor Day)
- License: Apache 2.0

List of abbreviations

ABCD standard: Access to Biological Collections Data standard; AWT: ARPHA Writing Tool; BOLD: Barcode of Life Data Systems; CSV file: comma-separated value file; DiSSCo: Distributed System of Scientific Collections; DwC: Darwin Core Standard; ENA: European Nucleotide Archive; FAIR data: Findable, Accessible, Interoperable and Re-usable data; GBIF: Global Biodiversity Information Facility; GGBN: Global Genome Biodiversity Network; GMI: Global Microbial Identifier; GSC: Genomics Standards Consortium; iDigBio: Integrated Digitized Biocollections; INSDC: International Nucleotide Sequence Database Collaboration; IPT: Integrated Publishing Toolkit; JATS: Journal Article Tag Suite; MIAME: Minimum Information About a Microarray Experiment; MINSEQE: Minimum Information about a high-throughput nucleotide SEQuencing Experiment; MIxS: Minimum Information about any (x) Sequence; MS: mass spectrometry; OBIS: Ocean Biogeographic Information System; OTU: operational taxonomic units

Competing interests

The authors declare that they have no competing interests.

Funding

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement IGNITE (No 764840) and from Pensoft Publishers.

References

1. Hey, T. and Trefethen, A., 2003. The Data Deluge: An e-Science Perspective. In: Grid Computing - Making the Global Infrastructure a RealityHey, A J G and Trefethen, A E (2003) . In, Berman, F, Fox, G C and Hey, A J G (eds.) Grid Computing - Making the Global Infrastructure a Reality. Wiley and Sons, pp.809-824.
2. Perez-Riverol, Y., Zorin, A., Dass, G., Vu, M., Xu, P., Glont, M., Vizcaíno, J., Jarnuczak, A., Petryszak, R., Ping, P. and Hermjakob, H., 2019. Quantifying the impact of public omics data. Nature Communications, 10(1).
3. *Darwin Tree Of Life*. 2020. Darwintreeoflife.org. https://www.darwintreeoflife.org/. Accessed on 8 June 2020.
4. *Earth BioGenome Project*. 2020. Earth Biogenome Project. https://www.earthbiogenome.org/. Accessed on 8 June 2020.
5. *Fondation Tara Océan*. 2020. Fondation Tara Océan. https://oceans.taraexpeditions.org/en/. Accessed on 8 June 2020.
6. Schriml, L., Chuvochina, M., Davies, N., Eloe-Fadrosh, E., Finn, R., Hugenholtz, P., Hunter, C., Hurwitz, B., Kyrpides, N., Meyer, F., Mizrachi, I., Sansone, S., Sutton, G., Tighe, S. and Walls, R., 2020. COVID-19 pandemic reveals the peril of ignoring metadata standards. Scientific Data, 7(1). https://doi.org/10.1038/s41597-020-0524-5
7. Karsch-Mizrachi, I., Takagi, T. and Cochrane, G., 2017. The international nucleotide sequence database collaboration. Nucleic Acids Research, 46(D1), pp.D48-D51.
8. Thessen, A. and Patterson, D., 2011. Data issues in the life sciences. ZooKeys, 150, pp.15-51.
9. *GBIF: The Global Biodiversity Information Facility*. 2020. What is GBIF?. https://www.gbif.org/what-is-gbif. Accessed on 25 June 2020.
10. *iDigBio*. 2020. https://www.idigbio.org/. Accessed on 25 June 2020.
11. DiSSCo. 2020. Home - Dissco. https://www.dissco.eu/. Accessed on 25 June 2020.
12. *OBIS (2020) Ocean Biodiversity Information System*. Intergovernmental Oceanographic Commission of UNESCO. www.iobis.org. Accessed on 25 June 2020.
13. Droege, G., Barker, K., Astrin, J., Bartels, P., Butler, C., Cantrill, D., Coddington, J., Forest, F., Gemeinholzer, B., Hobern, D., Mackenzie-Dodds, J., Ó Tuama, É., Petersen, G., Sanjur, O., Schindel, D. and Seberg, O., 2013. The Global Genome Biodiversity Network (GGBN) Data Portal. Nucleic Acids Research, 42(D1), pp.D607-D612.

14. Droege, G., Barker, K., Seberg, O., Coddington, J., Benson, E., Berendsohn, W., Bunk, B., Butler, C., Cawsey, E., Deck, J., Döring, M., Flemons, P., Gemeinholzer, B., Güntsch, A., Hollowell, T., Kelbert, P., Kostadinov, I., Kottmann, R., Lawlor, R., Lyal, C., Mackenzie-Dodds, J., Meyer, C., Mulcahy, D., Nussbeck, S., O'Tuama, É., Orrell, T., Petersen, G., Robertson, T., Söhngen, C., Whitacre, J., Wieczorek, J., Yilmaz, P., Zetzsche, H., Zhang, Y. and Zhou, X., 2016. The Global Genome Biodiversity Network (GGBN) Data Standard specification. Database, 2016, p.baw125.
15. Holetschek, J., Dröge, G., Güntsch, A. and Berendsohn, W., 2012. The ABCD of primary biodiversity data access. Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology, 146(4), pp.771-779.
16. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. and Vieglais, D., 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE, 7(1), p.e29715. https://doi.org/10.1371/journal.pone.0029715
17. *What Is Darwin Core, And Why Does It Matter?*. 2020. Gbif.org. https://www.gbif.org/darwin-core. Accessed on 15 July 2020.
18. Field, D., Amaral-Zettler, L., Cochrane, G., Cole, J., Dawyndt, P., Garrity, G., Gilbert, J., Glöckner, F., Hirschman, L., Karsch-Mizrachi, I., Klenk, H., Knight, R., Kottmann, R., Kyrpides, N., Meyer, F., San Gil, I., Sansone, S., Schriml, L., Sterk, P., Tatusova, T., Ussery, D., White, O. and Wooley, J., 2011. The Genomic Standards Consortium. PLoS Biology, 9(6), p.e1001088.
19. *Forside - Global Microbial Identifier*. 2020. Global Microbial Identifier. http://www.globalmicrobialidentifier.org/. Accessed on 15 July 2020.
20. Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J., Amaral-Zettler, L., Gilbert, J., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B., Blaser, M., Bonazzi, V., Booth, T., Bork, P., Bushman, F., Buttigieg, P., Chain, P., Charlson, E., Costello, E., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J., Gallery, R., Gevers, D., Gibbs, R., Gil, I., Gonzalez, A., Gordon, J., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A., Kelley, S., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C., Legg, T., Ley, R., Lozupone, C., Ludwig, W., Lyons, D., Maguire, E., Methé, B., Meyer, F., Muegge, B., Nakielny, S., Nelson, K., Nemergut, D., Neufeld, J., Newbold, L., Oliver, A., Pace, N., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D., Assunta-Sansone, S., Schloss, P., Schriml, L., Sinha, R., Smith, M., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J., Ward, D., Weinstock, G., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J., Yatsunenko, T. and Glöckner, F., 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology, 29(5), pp.415-420.

21. Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N., Petryszak, R., Papatheodorou, I., Sarkans, U. and Brazma, A., 2018. ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Research*, 47(D1), pp.D711-D715.
22. *Biosamples*. 2020. Ebi.ac.uk.https://www.ebi.ac.uk/biosamples/. Accessed on 21 May 2020.
23. *FGED: MINSEQE.* 2020. Fged.org. http://fged.org/projects/minseqe/. Accessed on 25 June 2020.
24. FORCE11. 2016. *The FAIR Data Principles*. https://www.force11.org/group/fairgroup/fairprinciples. Accessed on 12 June 2020.
25. Sansone, S., McQuilton, P., Rocca-Serra, P. et al. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol 37, 358–367 (2019). https://doi.org/10.1038/s41587-019-0080-8
26. Chavan, V. and Penev, L., 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics, 12(S15).
27. *Earth System Science Data*. 2020. http://www.earth-syst-sci-data.net/. Accessed on 7 July 2020.
28. Five years of Scientific Data. Sci Data 6, 72 (2019). https://doi.org/10.1038/s41597-019-0065-y
29. Penev L, Mietchen D, Chavan V, Hagedorn G, Remsen D, Smith V, Shotton D (2011). Pensoft Data Publishing Policies and Guidelines for Biodiversity Data. Pensoft Publishers, http://www.pensoft.net/J_FILES/Pensoft_Data_Publishing_Policies_and_Guidelines.pdf.
30. Penev, L., Mietchen, D., Chavan, V., Hagedorn, G., Smith, V., Shotton, D., Ó Tuama, É., Senderov, V., Georgiev, T., Stoev, P., Groom, Q., Remsen, D. and Edmunds, S., 2017. Strategies and guidelines for scholarly publishing of biodiversity data. Research Ideas and Outcomes, 3, p.e12431.
31. Penev L, Georgiev T, Geshev P, Demirov S, Senderov V, Kuzmova I, Kostadinova I, Peneva S, Stoev P (2017) ARPHA-BioDiv: A toolbox for scholarly publication and dissemination of biodiversity data based on the ARPHA Publishing Platform. Research Ideas and Outcomes 3: e13088. https://doi.org/10.3897/rio.3.e13088
32. *The ENA Metadata Model*. 2020. Ena-docs.readthedocs.io. https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.htm. Accessed on 5 May 2020.
33. Carpenter, E., Matasci, N., Ayyampalayam, S., Wu, S., Sun, J., Yu, J., Jimenez Vieira, F., Bowler, C., Dorrell, R., Gitzendanner, M., Li, L., Du, W., K. Ullrich, K., Wickett, N., Barkmann, T., Barker, M., Leebens-Mack, J. and Wong, G., 2019. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). GigaScience, 8(10).

34. Filho, J., Jorge, S., Kremer, F., de Oliveira, N., Campos, V., da Silva Pinto, L., Dellagostin, O., Feijó, R., de Menezes, F., de Sousa, O., Maggioni, R. and Marins, L., 2018. Complete genome sequence of native Bacillus cereus strains isolated from intestinal tract of the crab Ucides sp. Data in Brief, 16, pp.381-385.
35. Zhou, Y., Xiao, S., Lin, G., Chen, D., Cen, W., Xue, T., Liu, Z., Zhong, J., Chen, Y., Xiao, Y., Chen, J., Guo, Y., Chen, Y., Zhang, Y., Hu, X. and Huang, Z., 2019. Chromosome genome assembly and annotation of the yellowbelly pufferfish with PacBio and Hi-C sequencing data. Scientific Data, 6(1).
36. Programmatic Access To ENA Data. 2020. Ebi.ac.uk. https://www.ebi.ac.uk/ena/browse/programmatic-access. Accessed on 21 May 2020.
37. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2020). shiny: Web Application Framework for R. R package version 1.5.0. http://shiny.rstudio.com
38. RStudio, PBC, 2020. Shinyapps.io. Boston, Massachusetts, USA: RStudio, PBC.
39. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
40. Pensoft, 2020. Pensoft/Omics-Data-Paper-Shinyapp. GitHub. https://github.com/pensoft/omics-data-paper-shinyapp. Accessed on 18 August 2020.
41. Pensoft, 2020. *Omics Data Paper Generator*. https://mdmtrv.shinyapps.io/Omics_data_paper/. Accessed on 18 August 2020.
42. RStudio, PBC, 2020. Rstudio.Cloud. Boston, Massachusetts, USA: RStudio, PBC.
43. *Arrayexpress - Data Access Policy*. 2020. Ebi.ac.uk. https://www.ebi.ac.uk/arrayexpress/help/data_availability.html. Accessed on 5 May 2020.
44. National Information Standards Organization, 2019. ANSI/NISO Z39.96-2019, JATS: Journal Article Tag Suite, Version 1.2 | NISO Website. Niso.org. https://www.niso.org/publications/z3996-2019-jats. Accessed on 10 August 2020.
45. JATS4R, 2020. *JATS4R Validator*. Validator.jats4r.org. https://validator.jats4r.org/. Accessed on 18 August 2020.
46. *ESA's Ecological Archives*. 2020. Esapubs.org. http://www.esapubs.org/archive/default.htm. Accessed on 5 May 2020.
47. Smith, M., 2011. Data Papers in the Network Era. In: *Charleston Library Conference*. Against the Grain Press, LLC. http://dx.doi.org/10.5703/1288284314871. Accessed on 21 May 2020.

48. Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L. and Desmet, P., 2014. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLoS ONE*, 9(8), p.e102623.
49. *Data Papers*. 2020. Gbif.org. https://www.gbif.org/data-papers. Accessed on 5 May 2020.
50. Joachim Schöpfel, Dominic Farace, Hélène Prost, Antonella Zane. Data papers as a new form of knowledge organization in the field of research data. 12ème Colloque international d'ISKO-France :Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?, ISKO France, Oct 2019, Montpellier, France. halshs-02284548
51. Smith, V., Georgiev, T., Stoev, P., Biserkov, J., Miller, J., Livermore, L., Baker, E., Mietchen, D., Couvreur, T., Mueller, G., Dikow, T., Helgen, K., Frank, J., Agosti, D., Roberts, D. and Penev, L., 2013. Beyond dead trees: integrating the scientific process in the Biodiversity Data Journal. *Biodiversity Data Journal*, 1, p.e995.
52. Senderov, V., Georgiev, T. and Penev, L., 2016. Online direct import of specimen records into manuscripts and automatic creation of data papers from biological databases. *Research Ideas and Outcomes*, 2, p.e10617
53. Elie Arnaud, 2020. MetaShARK-v2. GitHub. https://github.com/earnaud/MetaShARK-v2. Accessed on 19 August 2020.
54. Susanna-Assunta Sansone, Philippe Rocca-Serra, Massimiliano Izzo, 2020. Datascriptor. https://datascriptor.org/. Accessed on 19 August 2020.
55. Martens, L. and Vizcaíno, J., 2017. A Golden Age for Working with Public Proteomics Data. *Trends in Biochemical Sciences*, 42(5), pp.333-341.
56. Filter, M., Candela, L., Guillier, L., Nauta, M., Georgiev, T., Stoev, P. and Penev, L., 2019. Open Science meets Food Modelling: Introducing the Food Modelling Journal (FMJ). *Food Modelling Journal*, 1.
57. Pensoft Publishers, 2020. *Data Quality Checklist And Recommendations*. Bdj.pensoft.net. https://bdj.pensoft.net/about#DataQualityChecklistandRecommendations. Accessed on 18 August 2020.
58. Kolesnikov N. et al., 2015. *ArrayExpress update-simplifying data submissions*. Nucleic Acids Res, doi:10.1093/nar/gku1057 . Pubmed ID 25361974.
59. Arrayexpress/Annotare2. GitHub. 2012. Web. Available at: https://github.com/arrayexpress/annotare2
60. Xue, H., Li, J., Xie, H. and Wang, Y., 2018. Review of Drug Repositioning Approaches and Resources. *International Journal of Biological Sciences*, 14(10), pp.1232-1244.
61. McDonald, D., Clemente, J., Kuczynski, J., Rideout, J., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. and Caporaso, J., 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(1).