WorldFAIR

Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D10.2 Agricultural Biodiversity Standards, Best Practices and Guidelines Recommendations

Work Package	WP10 - Agricultural Biodiversity
Lead Author (Org)	Debora Pignatari Drucker (Embrapa)
Contributing Author(s) (Org)	José Augusto Salim (Unicamp), Jorrit Poelen (Ronin Institute, UC Santa Barbara), Filipi Miranda Soares (USP & UTwente), Rocio Ana Gonzalez-Vaquero (UBA), Jeff Ollerton (UoN/KIB), Mariano Devoto (UBA), Max Rünzel (HiveTracks), Drew Robinson (HiveTracks), Muo

Kasina (KALRO), Christine Taliga (USDA NRCS), Cynthia Parr (USDA ARS), Diana Cox-Foster (USDA ARS), Elizabeth Hill (USDA OCS), Marcia Motta Maues (Embrapa), António Mauro Saraiva (USP), Kayna Agostini (UFSCAR), Luisa Gigante Carvalheiro (UFG), Pedro Bergamo (Unesp), Isabela Varassin (UFPR), Denise Araujo Alves (USP), Bruno Marques (UFG), Carla Tinoco (UFG), André Rodrigo Rech (UFVJM), Juliana Cardona-Duque (University CES), Mileidy Idárraga (University CES), M. Camila Agudelo-Zapata (University CES), Esteban Marentes Herrera (SiB Colombia), Maarten Trekels()

Due Date	31.01.2024
Date	19.01.2024
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.xxxxx

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	08.01.2024	Debora Pignatari Drucker	Initial draft
0.2	19.01.2024	All authors	Draft for internal review
			Content ready

Disclaimer

Abbreviations and Acronyms

CDIF	Cross-Domain Interoperability Framework
DwC	Darwin Core
FAIR	Findable, Accessible, Interoperable, Reusable
FIP	FAIR Implementation Profile
GBIF	Global Biodiversity Information Facility
GloBI	Global Biotic Interactions
IGAD	Improving Global Agricultural Data Community of Practice
IPBES	Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services
REBIPP	Brazilian Network of Plant-Pollinator Interactions
RDA	Research Data Alliance
TDWG	Biodiversity Information Standards
URI	Uniform Resource Identifier
USDA	United States Department of Agriculture

Executive Summary

The WorldFAIR Case Study on Agricultural Biodiversity addresses the challenges of advancing interoperability and mobilising plant-pollinator interactions data for reuse. Previous efforts, reported in Deliverable 10.1 - discovery phase - presented an overview of projects, good practices, tools, and examples for creating, managing and sharing data related to plant-pollinator interactions, along with a work plan for conducting pilot studies. This report presents results from the pilot phase of the Case Study, with six pilot studies that adopted standards and recommendations from the discovery phase. The pilots allowed for dealing with concrete examples and for the generation of reusable materials tailored to this domain, as well as for obtaining better estimates for the overall costs of adoption for future projects.

Our approach for plant-pollinator data standardisation is based on the widely-used standard for representing biodiversity data, Darwin Core, developed and maintained by the Biodiversity Information Standards (TDWG), allied to a data model and a vocabulary proposed by the Brazilian Network of Plant-Pollinator Interactions - REBIPP. The pilot studies also went through a process of FAIRification in the Global Biotic Interactions (GloBI, Poelen et al. 2014) platform. Moreover, we present the publishing model for Biotic Interactions developed together with the Global Biodiversity Information Facility (GBIF), which leads the Case Study on Biodiversity, as part of the proposed New Data Model, as well as a concrete example of its use by one of the pilots. This effort allowed for the development of 'FAIR best practices' guides for sharing plant-pollinator interaction data. This work primarily focuses on enhancing the interoperability of data on plant-pollinator interactions, envisioning its connection with the effort WorldFAIR is undertaking to develop a Cross-Domain Interoperability Framework.

We have successfully promoted the adoption of standards and increased the interoperability of plant-pollinator interactions data, resulting in a process that allows for tracing the provenance of the data, as well as facilitating the reuse of datasets which are extremely important to understand this essential ecosystem service and how it is changing due to human impact.

Our effort shows there are several possible paths for FAIRification, according to the institutional needs, and we demonstrated that different approaches can sum up to promote data interoperability and data availability for reuse, which is the ultimate goal of this initiative. As such, we have successfully achieved the goal of ensuring FAIR data for understanding plant-pollinator interactions at biologically-relevant scales for crops, with broad participation from initiatives in Europe, South America, Africa, North America and elsewhere. We also established concrete guidelines on FAIR data best practices customised for pollination data, metadata and other digital objects and promoted scalable adoption of these standards and FAIR data best practices by multiple initiatives. We believe this effort can help similar initiatives to adopt interoperability standards for this domain and contribute to our understanding of how plant-pollinator interactions sustain life on Earth.

Table of contents

1. Introduction

A huge amount of effort and attention has been given to scientific data in the last decades, with increasing intensity since the beginning of the millennium, as digital transformation in many fields and nations evolves. The FAIR principles were proposed a decade ago and became a worldwide reference for data management and stewardship. This topic will likely continue to gain importance with the accelerated development of data-driven research and development (e.g., Artificial Intelligence Models are trained on data) and data-integration challenges associated with the interwoven nature of the challenges currently faced by humanity. However, for many disciplines, much of the data generated by research projects around the globe are not reusable and, in many cases, there is still a gap between the need for trustable data and data availability for reuse. This is the case for plant-pollinator interaction data, as pointed out by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) Assessment Report on Pollinators, Pollination and Food Production (IPBES 2016). Since then, important efforts have been made towards standardising, integrating and publishing plant-pollinator interaction data, such as the Global Biotic Interactions (GloBI) platform, the Database of Pollinator Interactions (DoPI), and the data model proposed by the Brazilian Network of Plant-Pollinator Interactions (Salim et al. 2022).

Considering this scenario, the Research Data Alliance Improving Global Agricultural Data Community of Practice (RDA/IGAD CoP) identified initiatives related to plant-pollinator interaction data being carried out by different groups within the community. IGAD CoP assembled a representative team to collaborate on enabling FAIR data for this key ecosystem service vital for agriculture and many other mechanisms that sustain life on Earth.

The WorldFAIR Agricultural Biodiversity Case Study aims to establish concrete guidelines on FAIR data assessment and best practices customised for plant-pollinator interaction data, metadata and other digital objects, as well as to promote the application of the FAIR principles to plant-pollinator data. Following up on previous efforts undertaken during the discovery phase, described in Deliverable 10.1[1], "Agriculture-related pollinator data standards use cases report", this deliverable (10.2) presents results of the pilot phase of the Case Study, offering agricultural biodiversity standards, best practices and guidelines recommendations, which were developed considering lessons learned from pilot results. Deliverable 10.3 will complement this Case Study with agricultural biodiversity FAIR data assessment rubrics, as illustrated in Figure 1.

---

[1] Trekels, M., Pignatari Drucker, D., Salim, J. A., Ollerton, J., Poelen, J., Miranda Soares, F., Rünzel, M., Kasina, M., Groom, Q., & Devoto, M. (2023). WorldFAIR Project (D10.1) Agriculture-related pollinator data standards use cases report. Zenodo. https://doi.org/10.5281/zenodo.8176978.

Figure 1: Phases of the WorldFAIR Agricultural Biodiversity Case Study. Discovery Phase 10.1 produced Agriculture-related pollinator data use cases. Phase 10.2 (current phase) details existing selected pollinator data projects (or pilots) and generation of reusable materials for best practices. Deliverable 10.3 proposes methods to help review or assess usability (or FAIRness) of Pollinator Data.

During the pilot phase, we conducted pilot projects for standards adoption within the community, which allowed for the development of guidelines and recommendations, FAIR assessments and estimation of costs of adoption. Based on the workplan presented in D.10.1, the pilot's efforts allowed for promoting scalable adoption of these standards and FAIR data best practices by multiple and complementary initiatives and for producing concrete and comprehensive guidelines, which were lacking for standards for plant-pollinator data, metadata and other digital objects. The workplan was based in existing efforts as the Global Biotic Interactions platform (GloBI), the Brazilian Plant-Pollinator Interactions Network (REBIPP) data publishing model, the Plant-Pollinator interactions vocabulary (PPI) and the Global Biodiversity Information Facility (GBIF) New Data Model for Biotic Interactions.

In this report, we describe the work conducted with the pilot projects for standards adoption and the workflow used to adhere to the FAIR principles within the biotic interactions' domain, more specifically plant-pollinator interactions. We also present reusable materials tailored to this domain based on lessons learned by the pilots' results and estimate costs incurred for the adoption of standards. Moreover, we present the publishing model for Biotic Interactions developed together with the GBIF, which leads the Case Study on Biodiversity, as part of the proposed New Data Model, as well as a concrete example of its use by one of the pilots. Lastly, we describe how our efforts are aligned with the Cross-Domain Interoperability Framework (CDIF) and point to future developments.

2. Pilot Phase: Six Approaches to Agricultural Biodiversity Standards Adoption

2.1 Pilots Overview

Six different initiatives were selected to perform the pilots, with the requirement that the projects should be comprehensive of the diversity of plant-pollinator interaction data being generated in different regions of the world. The pilots are listed below and will be described in detail in the following sections.

2.1.1 Pilot - Observations of plant-pollinator interactions in the Pampean region of Argentina, conducted by Facultad de Agronomía - Universidad de Buenos Aires.

Five datasets comprising 120 interaction networks constructed from field observations documented across various locations in the Pampas, a region in central Argentina characterised by intensive cultivation (featuring soybean, maize, and wheat as its main crops), were converted to the Brazilian Plant-Pollinator Interactions Network (REBIPP) data publishing model, testing two different but comparable approaches. A tutorial was produced to facilitate similar efforts.

2.1.2 Pilot: The Brazilian Plant-Pollinator Interactions Network (REBIPP).

A collection of six datasets, encompassing data derived from distinct research projects on plant-pollinator interactions, was converted to the Brazilian Plant-Pollinator Interactions Network (REBIPP) data publishing model. The datasets are entitled "Orange (*Citrus sinensis* L. Osbeck, var. Pera-rio) insect floral visiting data of orchards in Itaberaí, Goiás, Brasil"; "Data compiled from published (original or review) studies carried out in Brazil on the reproductive system and pollination/pollinators of crop plants."; "Floral visitation in restored areas/remnants of natural vegetation in the Xingu region"; "Contribution of insect pollinators to orange production and quality"; "The Ecology of Plant Hummingbird Interactions (EPHI) - Brazil" and "Plant-flower visitor network from Avon Gorge, UK". The latter was also extracted to be used in the GBIF publishing model and was published in the GloBI portal.

2.1.3 Pilot - The Plant-Pollinator Interaction Data Collection by the Kenya Agricultural and Livestock Research Organization (KALRO) on the African continent.

A review of the status of African Plant-Pollinator Interaction data from various web-based sources such as journals, web pages, handbooks and manuals, which resulted in 1,030 records of animal-plant interactions. The data was converted into the GloBI simplified template and published in the GloBI portal.

2.1.4 Pilot- Pollinator Data from HiveTracks, a startup that works with smallholder beekeepers across the world to crowdsource environmental data collection.

HiveTracks provides their beekeeping mobile application to help beekeepers better track and understand their observations when visiting their apiaries, improve their management practices and the health of their bees. To support beekeepers in their goal, HiveTracks collects hive intervention, pollinator, plant, and interaction data that has been directly reported by beekeepers and is tied to specific geographic locations. The database for their current mobile application was mapped to both the Darwin Core (DwC) and BeeXML data standards.

2.1.5 Pilot - The United States Department of Agriculture Plant Pollinator Interaction Prototype Data and Database Development

The goal of the USDA pilot is to develop plant-pollinator interaction data tables as part of the USDA Natural Resources Conservation Service (NRCS) Plant List of Attributes, Names, Taxonomy, and Symbols (PLANTS) database in a way that will accommodate the desired data *and* be readily interoperable with other plant-pollinator datasets. The strategy for data standardisation was to design relational data tables based upon the recommendations of Salim *et al*. (2022), using Darwin Core, the Brazilian Plant-Pollinator Interaction Network (REBIPP) data publishing model, and Plant Pollinator Interactions Vocabulary (PPI) terms to define the database fields and to map existing fields from data sources. We created a prototype dataset drawn from two subsets of data. The first subset was data from the National Pollinating Insect Collection (NPIC) USDA-ARS Pollinating Insects Research Unit. The second is from the USGS Pollinator Library, USGS Northern Prairie Wildlife Research Center. Both sources include data from the peer-reviewed literature on known interactions in North America.

2.1.6 Pilot - Plant-pollination interactions in wild ecosystems, Colecciones Biológicas from the Universidad CES & SIB Colombia.

Universidad CES have been collaborating with SIB Colombia, the GBIF Node in the country, for some time and this group has experience on using DwC standard. The dataset "Web interactions between insects and some common plants in the "Refugio de Vida Silvestre Alto de San Miguel"", was built from a project which aimed on recording floral visitors in some common plants in a strategic area for Medellín city in Colombia and was standardised with DwC in the past. In this pilot, this dataset was converted to the REBIPP data model using information from photographs to allow for extracting the interactions details needed. The dataset "Pollination of the cycad *Zamia incognita* A. Lindstr. & Idárraga by *Pharaxonotha* beetles in the Magdalena Medio Valley, Colombia" came from the raw data gathered during a systematic study of the pollination system in a natural population of the endangered and endemic Colombian species of *Zamia* (published paper can be accessed at https://doi.org/10.1007/s11829-017-9511-y). After reviewing field notes and original data about thermogenesis, cones development and insect visitors, it was determined that this dataset could be a good example to test the versatility of relational tables of the new GBIF model. IDs for events, assertions, organisms, and material entities were created to link the data among the tables;

thereafter all the raw previously unpublished data was placed in each table. A new dataset was then created in the GBIF demo IPT for uploading the files.

2.2 Strategy for data standardisation

The pilot projects represent a diversity of data collection methods (e.g., literature review, direct observation, experiments in the field), geographic coverage (e.g., Africa, North America, South America, Europe, Middle East), digital data management methods (e.g., complex information system, pragmatic spreadsheet usage), and native language spoken (e.g., Portuguese, Spanish, English). Despite this variability, each pilot followed at least one of the recommendations described in Deliverable 10.1. More specifically, Darwin Core (DwC) was adopted as the main standard to describe generic aspects of biotic interactions data, and the Plant Pollinator Interactions Vocabulary (PPI)[2], a vocabulary of standardised terms maintained by the Brazilian Network of Plant-Pollinator Interactions (REBIPP), was used to document the specific details of plant-pollinator interactions data. The Ecological Metadata Language (EML) was used to standardise the metadata for each pilot, including licensing and other FAIR-relevant metadata such as contributor information, resource locations, and table definitions. The REBIPP data template proposed by Salim *et al.* (2022) was used to standardise the datasets.

To embrace the diversity of the pilot projects, we took the approach of facilitating the translation of existing data such that they could be compared and reviewed in a single framework. So, rather than mandating use of specific data formats up-front, pilot projects shared data in their native format.

Following, a collaborative approach was taken to transfer or convert the original data using a suite of templates developed by REBIPP (Salim 2023). This suite of REBIPP templates offered not only a familiar spreadsheet (tabular) format to express both metadata and the associated plant-pollination data records, but also offered a communication framework for our diverse group contributors of pilot projects. Where some pilot projects chose to make their data openly available for reuse, other pilots opted to restrict some, or all of their data products.

To keep track of the pilot projects and their associated (meta-)data, we reused the GloBI platform and integration methods. In these methods, each pilot project used a dedicated GitHub repository to manage metadata and register with GloBI (Figure 2).

---

[2] https://ppi.rebipp.org.br
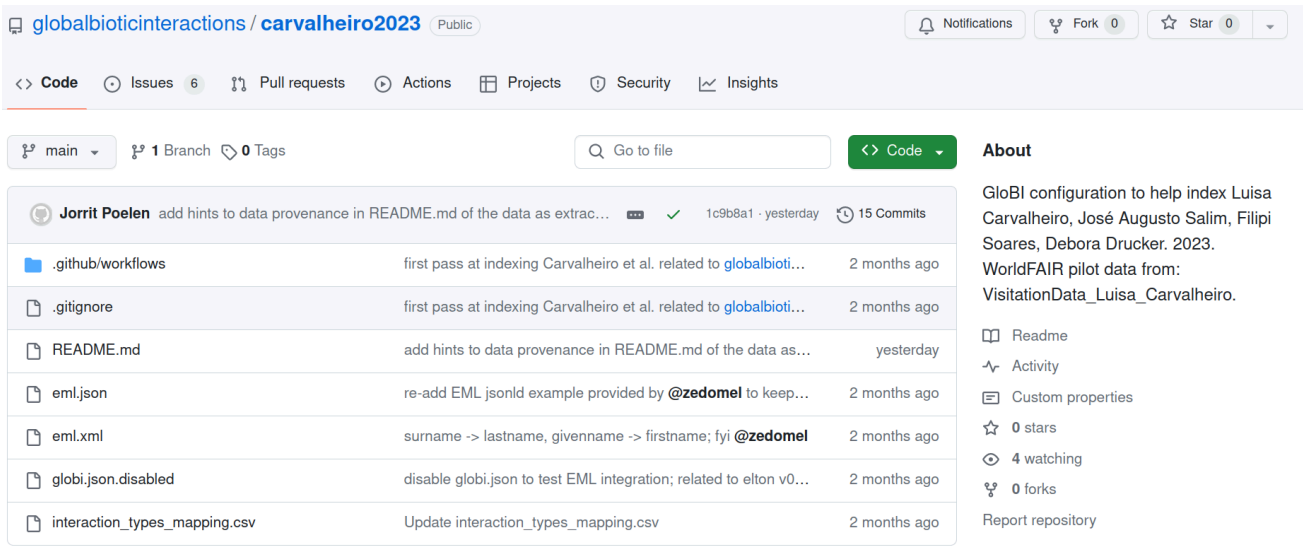
Figure 2. GitHub Webpage associated with the Carvalheiro et al. 2008 pilot as accessed via https://github.com/globalbioticinteractions/carvalheiro on 2023-01-17. The file "eml.xml" contains a machine readable version of the pilot metadata, including a reference to the actual pollinator data, hosted in Google Sheets. Also, the metadata contained the schema definition of the plant-pollinator data.

Figure 3. Automated GloBI review, search and data services

GloBI monitors federated (data) registries such as GitHub and Zenodo to find, access, integrate, and reuse species interaction datasets. As a result, derived data products are not only made available in multiple formats, such as RDF and CSV, but also through web-enabled data exploration services. This variety in data representation not only facilitates easier comprehension for human us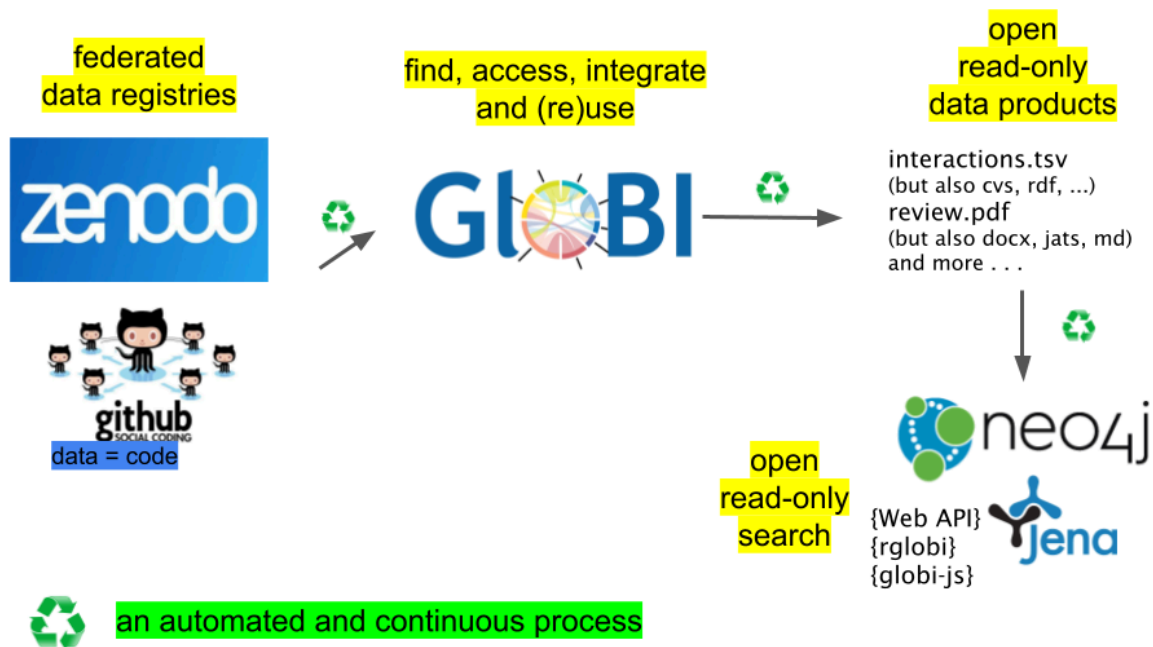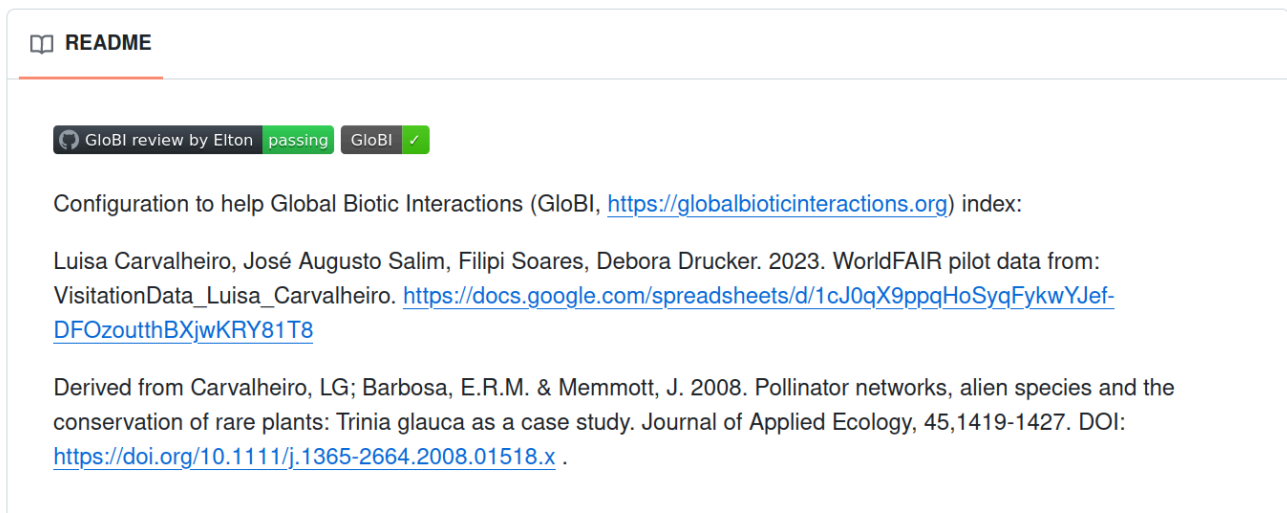ers but also ensures precise data reuse across different applications. So, by having the metadata associated with the Carvalheiro et al. 2008 pilot be independently published in a well-used platform (GitHub), a process of FAIRification is set in motion by GloBI services.

After creation of these pilot repositories, pilots are not only findable GitHub, general web search engines, but also benefit from automated GloBI review, search and data services (see Figure 3). Just like GitHub and web search engines index web resources, GloBI indexes web resources hosted on GitHub that mention "globalbioticinteractions" in their README.md (see Figure 4) and expose metadata in a machine-readable way that GloBI "understands" (e.g., through an eml.xml or globi.json file).

Figure 4. Content of README.md on Carvalheiro et al. 2008 pilot GitHub page as accessed https://github.com/globalbioticinteractions/carvalheiro on 2023-01-17. Note that the "GloBI review by Elton" and "GloBI" interactive badges indicate that the associated (meta-) data was reviewed and indexed by GloBI. This suggests that GloBI was able to find, access, integrate and reuse the Carvalheiro et al. 2008 pilot outcomes.

To help summarise and review the outcomes of our pilot projects, we created a dedicated WorldFAIR project page at https://globalbioticinteractions.org/worldfair. The webpage shows automatically generated pilot data review reports but also offers a way to manually inspect and review the outcomes of the various pilot projects and their associated contributors. See section 4.3 for a more detailed description of the WorldFAIR project page. Note that, in deliverable D10.3, these data review reports will be explored in detail as a method to assess the FAIRness of existing

plant-pollinator data. Just like we chose to reuse existing standards like EML, we leveraged existing infrastructures like GitHub and GloBI to keep track of, and summarise, the outcomes of our pilot projects.

Another key step taken by our approach was the implementation of controlled vocabularies in the process of data standardisation, which represents a fundamental step toward achieving semantic interoperability. Controlled vocabularies, which comprise a set of standardised and precisely defined terms, play a pivotal role in ensuring that data elements are consistently understood and interpreted across different systems and contexts. This standardisation is vital for semantic interoperability, the ability of different systems and organisations to not only exchange data but also to interpret and use the information meaningfully and accurately (Benson and Grieve, 2021; Turkmayali 2023). By utilising controlled vocabularies, organisations can overcome the challenges posed by variations in terminologies, ensuring that data elements retain their intended meaning across various platforms and interactions. This harmonisation of data language and meanings facilitates more effective and efficient data integration, exchange, and analysis, leading to improved communication and collaboration between diverse systems and stakeholders (Turkmayali 2023).

In this project, we utilised ontologies and standards such as DwC, PPI, EML, and the Relation Ontology (RO) to standardise terminology across datasets from various pilot studies in related fields. These studies often employ their own distinct terminologies, reflective of their unique contexts. To streamline this diversity and enhance clarity, we implemented a process known as data semantic annotation. Data semantic annotation is a critical procedure that involves mapping the natural language terms used in datasets with standardised terms from controlled vocabularies. This mapping is essential for maintaining data consistency and improving its interpretability.

Another approach we used in the pilot phase was developing the publishing model for biotic interactions of the new data model[3] in collaboration with GBIF. The publishing model serves as intermediate data representation to help data authors to standardise and publish their datasets through the IPT, GBIF, and other platforms that choose to adopt GBIF's new data model. A demo version of the Integrated Publishing Toolkit (IPT) with the new publishing model was tested by two of the pilot studies.

Our focus on collaboration and facilitation led to a common understanding that different projects have different needs, while also realising that work is needed to enable data integration between the various projects. In other words, our strategy was to keep local best practices in place while developing common ground to mobilise pollination data across our pilot projects as guided by FAIR principles and to use this experience as a basis for future projects that wish to standardise data in this domain. So, in our opinion, to FAIR-i-fy data means to enrich (or annotate, align, translate) existing datasets so that they can be combined with other datasets while meeting the needs of the

---

[3] https://www.gbif.org/new-data-model

diverse group of people and skills needed to help create, compile, share, review and curate the datasets.

3. Pilots Standard Adoption Stories

In this section, we present the standard adoption stories of each one of the six pilot projects. Each pilot report starts with an overview of the pilot study, followed by the data standardisation approach. We also present time investments to perform the pilot study and lessons learned by each adoption story.

3.1. Observations of Plant-pollinator interactions in the Pampean region of Argentina

3.1.1 - Overview of the pilot study

This pilot study focused on processing a substantial dataset comprising field observations of plant-pollinator interactions documented across various locations in the Pampas, a region in central Argentina characterised by intensive cultivation, primarily featuring soybean, maize, and wheat as its main crops. In highly managed agroecosystems, the significance of field margins in biodiversity conservation cannot be overstated, as these margins play a crucial role in supporting ecosystem services such as animal pollination and biological pest control, potentially benefiting crop yields.

For these reasons, the Facultad de Agronomía - Universidad de Buenos Aires has conducted numerous studies aimed at comprehending the intricacies of plant-pollinator interactions in this critical region for the country's productivity. Table 1 provides information about the datasets chosen for this pilot. Datasets are research projects (e.g., doctoral theses), whose sampling was performed in one to three spring-summer seasons. The sampling was carried out by postgraduate students, hired postdoctoral researchers, and field assistants during various field campaigns spanning from 2013 to the present.

A total of 120 interaction networks were constructed by sampling interactions between flowering plants and the insects that visited their flowers to forage for nectar or pollen. In cases where the adjacent crop featured flowers attracting insects (e.g., soybean, potato, and sunflower), sampling also encompassed flower visitors to the crop. An insect was classified as a visitor if it made contact with any of the floral sexual parts. Common and easily recognizable insect species were not collected; instead, their visits were meticulously recorded in a spreadsheet in the field. Plants and insects that could not be identified in the field were collected and later identified in the laboratory. Additional details about the methodology can be found in Monasterolo et al. (2020) and Tavella et al. (2022).

Table 1: Datasets processed for the pilot. The number of records represents the number of interactions (rows in a spreadsheet). FM = field margin. All counties belong to Buenos Aires province, Argentina.

Dataset	# networks	# records	Publication	Sampling	Year	County
1	20	1953	Monasterolo et al. 2020	FM	2013-15	Carlos Casares
2	29	518	unpublished	FM + soybean	2016	Carlos Casares
3	20	522	Tavella et al. 2022	FM + soybean	2018	Carlos Casares
4	35	1978	Thesis in progress	FM + potato	2020-23	Balcarce
5	16	781	Thesis in progress	FM + sunflower	2022-23	Balcarce

To implement this pilot, information regarding each dataset was available in one or more Excel spreadsheets. Each record (row) represents the interaction between an individual plant and a specific species of flower visitor (insect), with the number of flower visitor individuals recorded in a column labelled "frequency." Notably, the records lacked a unique ID, resulting in some rows being identical. Further details regarding most locations (site, latitude, longitude, elevation) were documented in another spreadsheet.

Datasets 1, 2, and 3 underwent prior curation by Dr Julia Tavella, who meticulously verified species identifications and amalgamated certain morphospecies (i.e. insects that could not be identified with certainty). Species not deemed potential pollinators (e.g., ants, predatory flies, leaf beetles) were excluded. These datasets comprised eight columns containing information pertaining to the interaction (campaign year, date, site), the plant (scientific name), and the animal (scientific name, family, order, number of individuals).

Datasets 4 and 5 belong to two doctoral theses in progress, and they include all flower visitors recorded during the surveys. These datasets comprised columns containing information pertaining to the interaction (date, site), the plant (scientific name, family, genus), and the animal (scientific name, order, family, genus, specific epithet, number of individuals). For dataset #5 in particular, the

field "scientificName" had to be constructed from the fields that contained information on the different taxonomic levels.

3.1.2 - Data standardisation approach

The process involved transforming the original spreadsheets into the final product: a REBIPP template filled with high-quality data, ready for upload and sharing on the REBIPP platform (https://db.rebipp.org.br/). All steps, including data cleaning, standardising, adding information, and performing taxonomic validation, were meticulously recorded. In this pilot, two methods were tested: a manual approach and a semi-automated approach. Datasets 1, 2, and 3 were selected to document each step required to standardise a spreadsheet and measure time investments, given that these datasets had been curated by the same person (Dr Julia Tavella) and shared the same format and number of columns.

Manual Approach:

In this method, single-function tools were employed, making it slower but less skill-intensive. Dataset 3 served as the basis for this approach. The cleaning and transformation of the data primarily occurred in Excel, utilising functions such as filters, "find and replace," and adding new columns with information. Canadensys date parsing (https://data.canadensys.net/tools/dates) and coordinate conversion (https://data.canadensys.net/tools/coordinates) tools were employed for standardising dates and coordinates, respectively. Taxonomic validation utilised the Global Names Resolver (https://resolver.globalnames.org), with the GBIF Backbone Taxonomy as the source. Species' authors and plant families were added during this process. When multiple possible authors were identified for a name, further research was conducted using GBIF Species search (https://www.gbif.org/species/search) and the Flora del Cono Sur catalogue (http://conosur.floraargentina.edu.ar/species/byscientificname) before making a decision.

Semi-automated Approach:

This method utilised a multi-function tool, making the process faster at the expense of requiring much greater expertise than that needed for the manual approach. Datasets 1 and 2 were used in this case approach. Data cleaning and transformation were performed using the OpenRefine program (https://openrefine.org/), a free, open-source tool that only requires Java JRE and an internet browser. OpenRefine allowed for mass transformation of data through facets, filtered views, and clustering to detect and merge alternative values. The program also facilitated matching datasets to external sources, such as Canadensys for automatic transformation of dates and coordinates. Taxonomic validation was facilitated by a routine that retrieved data from GBIF Backbone Taxonomy via GBIF's API. Authors and higher taxonomic levels were automatically obtained through this query. Visualising flagged names for revision in GBIF Species Matching (https://www.gbif.org/tools/species-lookup) made the validation process easier, as it sometimes

provided additional information. A step-by-step tutorial to apply this semi-automated approach can be found in Appendix 1.

Overall, these two approaches offer distinct advantages, with the manual method being accessible and the semi-automated method being efficient but requiring greater expertise.

3.1.3 - Time investments

The time allocated to each step of data standardisation can be viewed as "costs," categorised into two types for both approaches: fixed costs and variable costs.

Fixed Costs:

Fixed costs involve the time required to gain basic knowledge for correctly processing a dataset to be shared. These costs are acquired only once and are independent of the number of records or spreadsheets to be processed.

On one hand, acquiring knowledge about the Darwin Core standard and its terms involved watching a tutorial, reviewing terms in the Darwin Core Quick Reference Guide (https://dwc.tdwg.org/terms), and mapping spreadsheet columns to DwC terms. These tasks were completed in 540 minutes. The person handling the datasets for this pilot was already familiar with the DwC standard; otherwise, reading Wieczorek et al. (2012) would have been necessary.

On the other hand, learning the REBIPP terms included reading Salim et al. (2022), understanding the organisation of columns in the REBIPP template, and analysing which REBIPP terms can be completed for the datasets and are worth sharing. These tasks were carried out in 270 minutes.

An additional cost should be considered for the semi-automated approach: learning to use the program OpenRefine, needed to clean and standardise the datasets. This involved watching two tutorials (one each about basic and advanced functions), installing the program, and practising, following the guide written by Zermoglio et al. (2021). These tasks were carried out in 1380 minutes.

Variable Costs:

Variable costs included the time needed to clean the dataset, add new columns with shareable information, and standardise the data. Taxonomic validation of names, for both plants and animals, and transferring the columns to the REBIPP template and completing the metadata section are also variable costs.

These costs depend largely on how clean and complete a spreadsheet is, as well as the number of records and scientific names it contains. For instance, if a spreadsheet has dates and coordinates in different formats for each record, cleaning this data will take much longer than if all records have the same format and values can be easily transformed into the standard format en masse. Similarly,

a spreadsheet with many scientific names with typographical errors will complicate the validation process, extending the time required.

To share a dataset in REBIPP, completing the REBIPP template needs to be considered a variable cost. This involves transferring columns to the REBIPP template, dependent on how many columns will be shared, and completing the metadata section. This time does not depend on the number of records or species included in a given dataset. For both approaches, a time of 300 minutes was estimated to carry out these tasks and check all the information before sharing it on the REBIPP platform.

Table 2 summarises the time invested in each task regarding fixed and variable costs for both approaches. Considering the variable costs, the process of cleaning, adding columns with new information, and standardising the data to the DwC/REBIPP format was much faster when applying the semi-automated approach (0.35 minutes/record) than the manual approach (1.70 minutes/record). Additionally, taxonomic validation was faster when performed through a routine in OpenRefine (4.30 minutes/name) compared to consulting Global Names Resolver and modifying the dataset manually in Excel (5.07 minutes/name). Although this difference in time was not so pronounced, the OpenRefine routine automatically retrieves the authors and higher taxonomic levels of the species, completing these columns for all records. For this reason, the semi-automated approach, in addition to being faster, is less prone to human errors.

Table 2: Time (in minutes) required to perform each task in the manual and semi-automated approaches. The number of names for validation does not include morphospecies identified by numbers. The number of records represents the number of interactions (rows in a spreadsheet) processed under each approach (for more details, see section 3.1.1). NA = not applicable.

	Manual	Semi-automated
Learn DwC	540	540
Learn REBIPP	270	270
Learn OpenRefine	NA	1380
∑fix costs	810	2190
Cleaning + standardization	890	870
Taxonomy validation	340	950
REBIPP template	300	300
∑variable costs	1530	2120

#names for validation	67	221
#records	522	2471
∑variable costs/#records	2.93	0.86

The time invested in each approach (Table II) is graphed in Figure 5, where the sum of the fixed costs is the y-intercept, and the slope is the time per record.
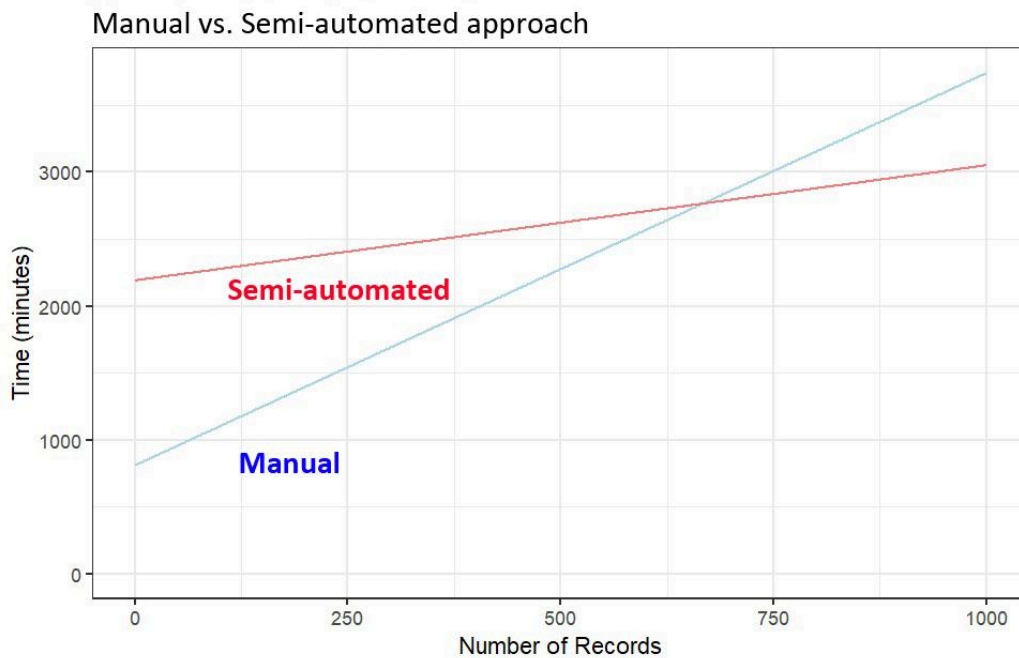

Figure 5. Functions for the manual and semi-automated approach. Manual approach function: Time = 2.93 x #records + 810. Semi-automated approach function: Time = 0.86 x #records + 2190. The intersection of the functions occurs when the number of records is 666.

3.1.4 Lessons learned and recommendations

Planning

Prior to embarking on a standardisation project, careful planning is crucial. Considerations should include the selection of datasets for sharing, identifying responsible individuals to carry out the

work, and providing a realistic estimation of time investment. Creating a comprehensive plan and avoiding underestimating the time required for each step is strongly recommended. Notably, the time investments presented here assume prior knowledge about the origin of the data, essential when the person standardising the dataset is not the original sampler. In this pilot, spreadsheets prepared by others were processed, necessitating contact with the preparers for clarifications and missing information.

Questions to be asked during planning include the dataset's cleanliness, the extent of missing information (dates, coordinates, collectors), responsibility for sampling, data reliability, and the availability of the preparers for further information. Additionally, it is essential to determine whether the data have been published or are part of an ongoing project and to ensure the necessary permissions for public sharing. The context in which the standardisation process will occur is also important; a dedicated person will likely result in a more efficient process.

The background of the person conducting the work is vital. A profile in biological sciences/biodiversity with knowledge of botanical and zoological nomenclature is necessary. Basic skills in Excel and database management are required. Familiarity with the species in the dataset facilitates taxonomic validation, and a background in plant-pollinator interactions and plant reproductive biology aids in interpreting REBIPP terms. Basic knowledge of georeferencing may be useful to understand some DwC terms from the Location class. Managing JSON (JavaScript Object Notation) file format may also be beneficial, to elucidate information retrieved in OpenRefine.

Manual or Semi-automated Approach?

The choice between a manual or semi-automated approach depends on the skills of the person in charge and the complexity of the data. Fig. 1 provides a helpful approximation, suggesting that for fewer records, the manual approach is preferable. However, when dealing with a relatively large number of records (e.g., for this pilot, over 666), the semi-automated approach becomes more efficient, particularly for future datasets or large datasets.

While the time investment per record may initially be higher for the semi-automated approach, it decreases significantly after the first spreadsheet is processed. Familiarity with processing information and species names in the same biogeographic region contributes to this reduction. This holds true for both methods but is more pronounced for the semi-automated approach, as OpenRefine allows saving of workflows and applying steps performed on one dataset to another in the future (see tutorial in the Appendix).

3.2. The Brazilian Plant-Pollinator Interactions Network (REBIPP).

3.2.1 - Overview of the pilot study

The Brazilian Network on Plant-Pollinator Interaction (known as REBIPP, accessible at https://www.rebipp.org.br) is a collaborative platform of experts in Pollination Biology. Its primary focus is the comprehensive study of plant-pollinator interactions across diverse dimensions, aiming to foster the advancement of scientific knowledge and educational initiatives in this field.

In this pilot initiative, we invited REBIPP members to contribute with their plant-pollinator interaction datasets for data "FAIRification". A total of six datasets were considered during this pilot phase, encompassing data derived from distinct studies focusing on plant-pollinator interactions.

The original datasets underwent transformation using the specialized REBIPP template, available at http://db.rebipp.org.br/how-to-contribute, thereby conforming to the Darwin Core standard. Furthermore, metadata information was created using the EML standard. Both the data, when openly available, and the corresponding metadata were subsequently published in GloBI.

3.2.2 Data standardisation approach

Data standardisation began with the migration of data from individual datasets to the corresponding fields within the REBIPP template. However, due to mismatches between some columns/fields in the original data and the Darwin Core (DwC) standard and the Plant-Pollinator Interactions Vocabulary (available at https://ppi.rebipp.org.br), these fields were not integrated into the REBIPP template. Most of them represent knowledge instead of evidence, and they were included by the authors to facilitate their analysis. However, some of them were not included due to the lack of terms or context, such as pollinator efficiency, indicating a concept that needs to be considered for being incorporated to the PPI vocabulary; or some chemical properties of fruits, an opportunity to practise cross-domain interoperability within WorldFAIR and that is being explored with the Case Study in Chemistry.

We engaged data authors in providing metadata descriptions directly within the spreadsheet housing the data. These metadata descriptions were useful for generating standardised metadata files conforming to the Ecological Metadata Language (EML).

We did not perform any taxonomic validation or data quality checks of the datasets. Our primary focus during this pilot was to assess the feasibility and implications of adopting FAIR principles, particularly centred around (meta)data standardisation.

3.2.3 Time investments

During the data standardisation of processing of each of the dataset, we individually measured the time spent on these tasks to estimate the efforts for the adoption of this approach for data standardisation. Table 1 shows the total time for each dataset standardisation, separating the time spent in data extraction, creation of GitHub repositories to host the datasets, and setup of the dataset for being indexed by GloBI (publishing). Table 3 also includes the total number of records N in each dataset as well as the estimate of the time spent on the standardisation of each record.

Table 3. Total amount of time spent in initial standardisation and publication of REBIPP datasets.

Dataset	Extraction	Data repositories	Publishing	Total time	Records/hour
Carvalheiro2023 (N=685)	03:52:40	0:05:00	02:30:00	06:27:40	106.50
Bergamo2020 (N=1588)	14:56:12	00:33:12	00:12:02	15:41:26	101.21
Ferreira2023 (N=58)	01:01:19	00:21:14	00:15:58	01:38:31	35.32
Alves2023 (N=48)	03:00:27	00:20:25	00:09:46	03:30:38	13.67
Tinoco2023 (N=515)	05:40:17	00:22:43	00:11:05	06:14:05	82.60
varassin2023 (N=22248)	03:54:09	00:15:58	00:13:19	04:23:26	5067.24
Average time	**05:24:11**	**00:22:42**	**00:12:26**	**05:53:28**	**913.04**

3.2.4 Lessons learned and recommendations

The process of data transformation can prove to be both time-consuming and costly, often significantly influenced by the original data's format and structure. During the data transformation of REBIPP datasets, we have observed an unexpected trend: larger datasets tend to be more straightforward to standardise compared to their smaller counterparts. For instance, the 'varassin2023' dataset, comprising 22,248 records, took approximately 3 hours and 54 minutes to be extracted. This specific dataset organises interaction records in a row-wise fashion, where each interaction between a plant and a floral visitor is condensed into a single row. Given this layout, mapping columns from the original data to Darwin Core and PPI fields becomes a relatively straightforward process, significantly influenced by the total record count.

Conversely, the 'alves2023' dataset, consisting of only 48 records, posed a challenge and needed a transformation time of 3 hours. Unlike the previous dataset, 'alves2023' doesn't organise interaction records row-wise. Instead, it adopts a complex matrix structure, arranging plant occurrences along rows and floral visitors across columns, accompanied by additional attributes in separate columns. This matrix structure is commonly used in ecological network studies (Salim et al. 2022). The transformation from this representation to the event (row-wise) representation requires replicating the plant occurrences for each floral visitor in the columns. This means that the original dataset with 16 rows representing the plant occurrences and 36 columns of floral visitors would result into a maximum of 16*36 (576) records in the event representation, but usually the final number of records is smaller than this, because for some columns the number of visitors is zero (no visitor). This particular organisation, aligned with the data model for biotic interactions, treating interactions as records composed of a source organism interacting with a target organism, demanded a more intricate transformation process extending beyond a simple column mapping. The transformed data has 48 interaction records in row-wise format.

Therefore, initiating data recording in a structure that facilitates subsequent transformation becomes a key aspect of data digitization and standardisation. Researchers and data authors must recognize the significance of structuring their data effectively. Doing so not only minimises the need for unwarranted and unforeseen efforts during subsequent data standardisation, but also greatly facilitates data sharing.

3.3. Plant-Pollinator Interaction Data Collection by the Kenya Agricultural and Livestock Research Organization (KALRO) on the African continent.

3.3.1 - Overview of the pilot study

A review of the status of web-based African Plant-Pollinator Interaction data was conducted by KALRO. In the context of this activity was the information about the interaction of a reported pollinator and reported plant, and other information about such interaction, including, but not limited to, the place where interaction occurred, and reference material which reported the interaction. A protocol was developed to collect and mine data from various web-based sources such as journals, web pages, handbooks, manuals and any other source that is web accessible and has the information sought. The intention was to collect as much as possible data on plant-pollinator interactions from studies carried out in Kenya and other parts of Africa.

3.3.2 Data standardisation approach

To assure technical support and homogeneity of approaches, the activity was carried out in a workshop setup whereby a team of data miners was brought together for 5 days to collect the data. They were trained on the meaning of plant-pollinator interactions, keywords that should be looked for, data coverage area, specific sites that are useful, and the protocol for reporting. Further,

monitoring and review of the data collection was continuously done, to ensure the right data was collected appropriately.

Various freely available internet based search engines such as Google Scholar were used to collect the data. The focus was mainly the internet-accessible/available data. The data sources were from, Journal articles and academic theses e.g. MSc, PhD Books, handbooks, manuals, technical reports, brochures, and articles (scientific, news, blogs) in websites.

The data was registered with the GloBI platform through reuse of their dataset template to enhance its use based on the FAIR data principles. The purpose of the copy-paste-edit GloBI dataset template is to facilitate exchange of existing species interaction datasets like KALRO. The GloBI dataset template can be found at https://github.com/globalbioticinteractions/template-dataset and contains instructions for use. Note that the template represents only one of many ways to register data with GloBI.

3.3.3 Time investments

The first exercise involved data collection by Information Communication Technology (ICT) analysts who have experience in data mining. A massive amount of data was collected using the available data mining tools. However, the review process only approved less than 30% of the mined data and the rest was rejected because it had nothing to do with the pollinator interactions.

Another episode of data mining was organised where biologists (BSC, Msc, PhD) were involved in data mining and data entry. Further, they were trained on the keywords and search processes. They were also trained on the GloBI requirements. They were then allowed to carry out the data collection and entries.

To enhance productivity, meeting workshops were held to bring project participants together to enhance the data collection. Initially, a team of 10 persons drawn from Information Communication Technology Division KALRO and JKUAT, was brought together in one week, investing on average 8 hours per person per day (total about 40 hours) to collect the data. In total we can estimate 400 hours by 10 persons to produce about 600 database entries, which translated to 120 entries after data cleaning. An entry is a row of information entered into the database. The information is entered based on the database requirements. In some instances, the entry may not completely provide every requested detail, since this is based on the source of the information. Thus from this exercise, a clean and relatively good database entry will take about 3.3 hours to enter.

Data collection by experts is more time-saving, with an average of about 30 minutes to fully complete an entry. A further 10 minutes is spent to validate the entry by the experts.

3.3.4 Lessons learned

Data collection by data specialists is fast because of using the mining tools. However, it needs to be well curated because it may not present what is needed. Data mining by biologists requires less

vigorous confirmations because the expertise picks close to what is required. Data mining by biologists is true to type, implying that the right data is collected with no or minimal errors. Generally, biologists have a better comprehension of plant-pollinator interactions. However, they still need validation to clean the data from the pollinator experts. However, these specialists are few and more time is required to engage them to collect enough representative data

Non pollinator-plant interaction experts have challenges entering the data. This is because it is not the area of their expertise. Search engines will only provide data results based on the key words used. The results can be so large, requiring expert knowledge to provide the correct information.

In instances where data is paywalled, it is impossible to mine the required information because some of those articles do not explicitly provide plant-pollinator interaction information from the public accessible sites. For example, an article providing information about a group of taxa in relation to interaction with plants may provide a summary of interaction at groups level. However, in our reporting, we are more interested with species level, which may be accessed after bypassing the paywall.

3.4. Pollinator Data from HiveTracks

3.4.1 Overview of the pilot study

HiveTracks provides a free beekeeping mobile application to help smallholders beekeepers better track and understand their observations when visiting their apiaries to increase bee health and monitor the environment (i.e., iOS and Android). Since 2010, HiveTracks has been used by over 40,000 beekeepers across 150 countries. To support beekeepers, the HiveTracks App helps collect hive intervention, pollinator, plant, and interaction data that is directly reported by beekeepers and is tied to specific geographic locations. Specifically, through HiveTracks' record system, occurrences of animals, plants, or interactions between two or more actors, as well as instances of beekeeper activities within an apiary, are recorded for a specific time and location in HiveTracks' remote database.

The HiveTracks app is designed to guide beekeepers to collect relevant and standardised data points regardless of their level of technical expertise (i.e., compare Figure 6). In combination with its global reach, this has the potential to achieve a continuous global monitoring of bee and apiary related data. In addition, this data collection process allows HiveTracks to aggregate beekeeping/pollinator data within and across hives and apiaries to understand trends ranging from the health of one hive to the average honey harvest of every apiary within a given region. The goal of this pilot study was to identify relevant data standards beyond the Darwin Core standard that could be relevant to the beekeeping community and map example data to these standards to lay the groundwork for an automated mapping. Furthermore, the goal of this pilot study was to show a pathway toward deeper private sector involvement in the topic of data standardisation.
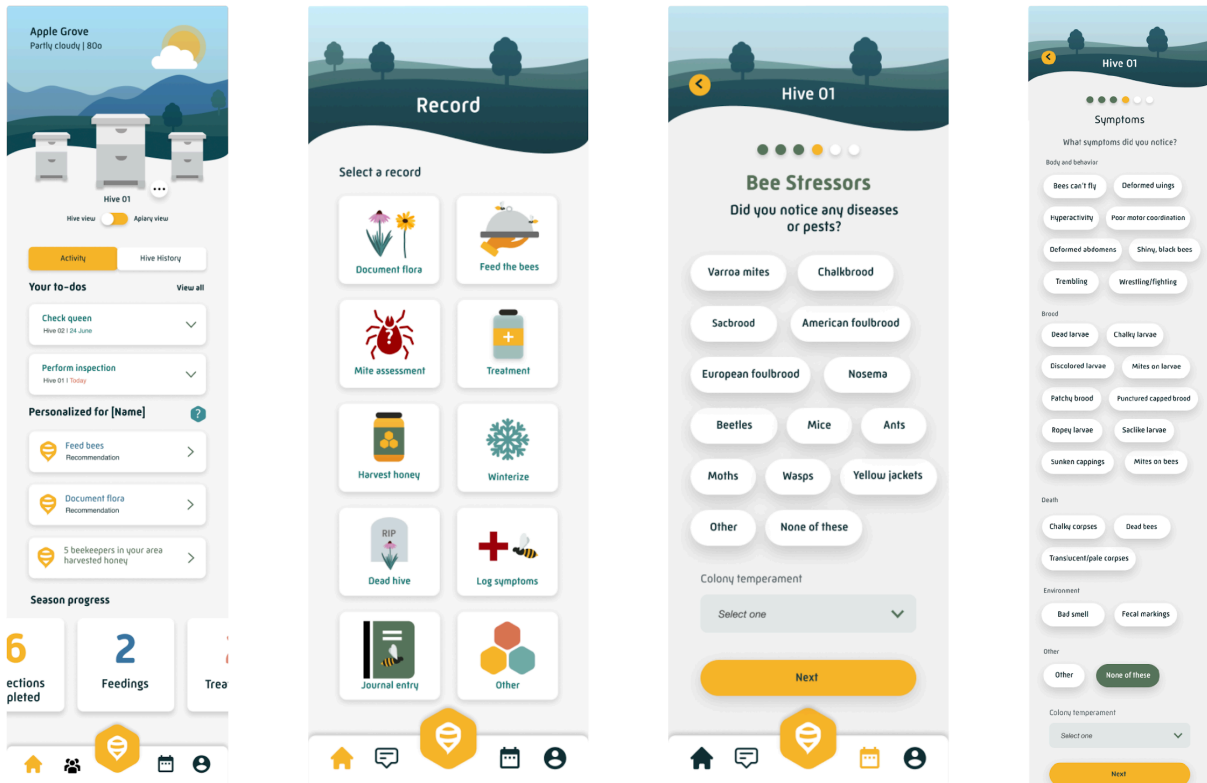
Figure 6: Screenshots from HiveTracks Mobile App showing home, record selection, bee stressor, and symptom screens

3.4.2 Data standardisation approach

The HiveTracks team focused on mapping the database for their mobile application to both the Darwin Core (DwC) and BeeXML data standards, the latter was identified as a relevant standard for the beekeeping community. For context, the Apimondia Bee XML working group describes BeeXML as "a self-describing data format that allows the exchange of data on bees and beekeepers" (https://beexml.org/). These mappings demonstrate the opportunity for HiveTracks' app to provide an ongoing source of new pollinator, plant, and pollinator stressor/pollinator interaction occurrence data (DwC) along with data on the specific actions being performed by the beekeeper within their apiary, such as a honey harvest (BeeXML). In addition, the HiveTracks team showed how certain tables/fields in their database could automatically be mapped to these existing data schemas for efficient individual and aggregate (with other data sets) analysis of HiveTracks' data. In sum, the HiveTracks team aimed to understand two existing data standards (DwC and BeeXML) and then

analyse the existing (and potential) alignment between these standards and the data HiveTracks is collecting through their mobile application for standardisation.

3.4.3 Time investments

The HiveTracks team spent 13.5 hours researching, developing, discussing, and presenting the data mappings to Darwin Core and BeeXML standards. Specific tasks included:

- (i) Analysing the work-in-progress REBIPP data template for DwC field - referring to the official published Darwin Core quick reference guide when further clarification was required.
- (ii) Analysing the existing BeeXML standards, which are being revised by the BeeXML team and have not been released at the time for writing.
- (iii) Identifying mappings between these two standards and the HiveTracks database schema - internal WP10 meetings were conducted for feedback on the DwC mapping to ensure the HiveTracks team's understanding of the Darwin Core standard was sound. For this pilot, plant mappings were particularly explored here due to HiveTracks' implementation of the Pl@ntNet API (https://my.plantnet.org/) allowing for the automatic classification of plants based on a photo submitted through the mobile app. Specific mappings for stressor interactions within DwC were also mapped out by the HiveTracks team based on interaction type mappings provided by iNaturalist (https://www.inaturalist.org/).
- (iv) Querying the HiveTracks database to provide sample data (simulated by HiveTracks team, not actual, real-world user data, to ensure proper data protection and privacy for HiveTracks users) for all mappings. This included data from tables pertaining to the beekeepers, apiaries, hives, colonies, queens, and records along with the metadata HiveTracks collects on each of these entities through HiveTracks' mobile application. These data were included in the mapping files with both the mappings and sample data explicitly pointing to one another for clarity.

3.4.4 Lessons learned and recommendations

Lessons learned: For the DwC standard, HiveTracks' data can cover multiple plant, animal, and interaction occurrences. The data schema is especially well-suited to capture stressor-to-pollinator interactions (e.g., Varroa mites, wasps, beetles, ants, moths, etc.; compare the two pictures to the right in Figure 6), occurrences of bee populations, including genetics, and occurrences of plants (i.e., the Pl@ntNet API can predict the plant species from a picture). For the current BeeXML standard, HiveTracks' data maps to key hive event and transaction data. However, as this standard will be updated significantly a revision of the mapping will have to be carried out. Overall, the pilot was very helpful to better understand the potential applications for the data and, in turn, how HiveTracks can better collect and organise these data for these applications, paving the way for real-time mapping of the HiveTracks data to the DwC and the forthcoming BeeXML standard. There are limitations to strict adoption of specific standards since different standards focus on different aspects of the data (i.e., DwC comparted to BeeXML) that may not optimally align with user / beekeeper requirements. For example, the current edition of BeeXML requires standardised metric

units for any event with an intervention that can be measured. However, allowing users to enter a free text response for units, in turn, allows for unconventional yet accurate measurement options relating to specific beekeeping tools and products (without users translating these measurements into potentially inaccurate standard units). Nonetheless, awareness of these various standards, especially with regards to FAIRness, has been beneficial to assessing the state of HiveTracks' data pipeline and how it can be improved going forward to adhere to the current standards.

Recommendations: As for continuing the standardisation work, monitoring the new BeeXML release that's expected to come within the next six months to a year will be essential. Once released, this new schema should be incorporated into the work-in-progress REBIPP data template in addition to finding the relevant mappings with existing data sets. BeeXML also accounts for extensions into other schemas within their schema (in the current iteration), so this should be considered as a potential avenue for efficiently addressing multiple data standards at once. Also, additional research and similar analysis techniques should be performed for other data standards that emerge in the future addressing beekeeping, pollinator, and/or biodiversity data. Furthermore, HiveTracks should use this pilot study to demonstrate the benefits for private sector companies to engage in data standardisation efforts. Finally, to test the real-time mapping of its database to the DwC and BeeXML, additional pilots should be carried out in the future.

3.5. The USDA Plant Pollinator Interaction Prototype Data

3.5.1 Overview of the pilot study

The United States Department of Agriculture (USDA) provides leadership on food, agriculture, natural resources, rural development, nutrition, and related issues and to preserve natural resources through conservation, restored forests, improved watersheds, and healthy private working lands. Given the paramount importance of pollinators to the health of agriculture and natural resources, the USDA Office of the Chief Scientist (OCS) is coordinating Honey Bee and Pollinator Research across multiple, diverse USDA mission areas and fellow federal agencies to ensure harmonised and successful coordination of resources. As part of this work, USDA is interested in promoting cross-agency collaborative efforts to share available pollinator data that is generated, funded, or collected in collaboration with USDA. An emphasis is placed on satisfying FAIR data principles to ensure these data are Findable, Accessible, Interoperable, and Reusable to increase government-to-citizen communication and transparency. To further these goals OCS is working with representatives from the Research Education and Economics (REE) mission area, Agricultural Research Service (ARS), and Natural Resources Conservation Service (NRCS) to pilot initiatives for inter- and intra-agency data sharing. One such pilot effort is to evaluate data sharing among the ARS National Pollinating Insect Collection (NPIC) and the NRCS PLANTS (Plant List of Attributes, Names, Taxonomy, and Symbols) Database. The NPIC is a world-class collection of bees and related wasps supporting research on pollination, systematics, biodiversity, and conservation. The database includes approximately 2.3 million specimens of which approximately 25% include plant pollinator interaction data, covering 137 nations worldwide with an emphasis on the US. NPIC

data is already indexed by GBIF (https://scientific-collections.gbif.org/collection/c0e5f597-c80b-421c-acb2-c5b559e73f46), but the indexed data does not yet include plant-pollinator interactions even when they are present in the NPIC database. The PLANTS database serves as the national standard for plant taxonomy and provides information on plant characteristics and spatial distribution of over 35,000 species of plants naturally occurring in the United States, sovereign Nation Lands, territories and protectorates. With over 12.5 million global page views annually, the PLANTS database and website (https://plants.sc.egov.usda.gov/home) is one of the most widely utilised websites in USDA. The PLANTS database is expanding its scope to include pollinator interaction data associated with specific plant species. This pilot includes an evaluation of the data standards to be integrated into the data framework within the newly developed pollinator interaction tables of the PLANTS database. This data will comprise existing publicly available USDA-funded plant-pollinator interaction research data (including data from NPIC) as well as data extracted from a systematised review of scientific peer-reviewed literature of plant-pollinator interactions within the US, territories, and protectorates. Of course, in addition to adding pollinator information to the enriched PLANTS public-facing web interface, the ultimate goal is to ensure that these data are interoperable when shared on an ongoing basis with similar global datasets.

3.5.2 Data standardisation approach

The strategy utilised for data standardisation was to:

1) Research and refinement. Review Darwin Core biodiversity information standards and the REBIPP template in order to create a mapping spreadsheet of the Darwin Core and REBIPP fields to existing fields in the data sources, NPIC and the planned pollinator interaction enhancements to the PLANTS database.

2) Data framework and organisation. Design a schema for the new pollinator interaction tables of the PLANTS database that will accommodate the desired data AND be readily interoperable with other plant-pollinator datasets.

3) Iteratively refine the data framework. Discussions among OCS, REE, ARS, NRCS regarding NPIC data and PLANTS pollinator data framework.

4) Assemble sample data from the sources, assigning column headers drawn from Darwin Core standard and REBIPP PPI vocabulary. These sample data were then submitted for evaluation.

In this pilot we had existing partial datasets, but the source data were not yet fully integrated into one system. Therefore, the pilot required both new database design as well as a prototype for sharing the data from the database. Although the prototype data had only 196 rows, this paves the way for future large scale data integration and sharing. Considering the standards up front during the database design phase should greatly facilitate interoperability among data sets when the system is fully developed.

3.5.3 Time investments

Adopting the data standardisation approach was initially time intensive as the USDA pilot participants became familiar with these concepts. The following is a breakdown of time by specific tasks:

• Darwin Core research and refinement. This time was spent by all USDA participants, from OCS, REE, ARS, NRCS.   80 hours

 • Data framework and organisation by the NPIC and PLANTS data managers. 80 hours

  • Discussions among all USDA participants regarding the pilot, NPIC data, and PLANTS pollinator data framework. 25 hours

 • Preparing datasets and discussion with the WorldFAIR support team. 40 hours

Total time approximately 225 hours

3.5.4 Lessons learned and recommendations

Leveraging the Darwin Core biodiversity standards is time-consuming for researchers who are unfamiliar with the standards. The REBIPP template is extremely helpful and provides a significant time-saving strategy. Allowing time for research and training to facilitate incorporating the interoperability standards within the initial framework of the data should provide for significant time savings. This pilot has facilitated the sharing of information to foster common learning and synergies. This process has been helpful for inter- and intra-departmental pollinator efforts and sets the stage for multiple data integration and sharing efforts across USDA and with partners such as the US National Bee Monitoring Research Coordination Network (RCN).

3.6. Plant-pollination interaction in wild ecosystems, Colecciones Biológicas from the Universidad CES + SIB Colombia

3.6.1 Overview of the pilot study

After the Embrapa presentation during the meeting "New data model: Exploring interactions data on plant pollination"[4], our group at the University CES asked for the possibility of testing the new model with some of our heterogeneous projects on insect-plant interactions. Part of the data came from systematic studies for describing wild pollination systems and the other part came from efforts for recording flower-insect interactions. In this pilot two datasets were considered: *1)* we selected a small dataset from field observations (and photographic records) which were previously adapted to

---

[4] https://www.gbif.org/composition/6yJQMg2cvnnzb88xwt62m7/exploring-interactions-data-on-plant-pollination

a standard DwC template called "Web interactions between insects and some common plants in the "Refugio de Vida Silvestre Alto de San Miguel"" (San Miguel dataset), built from a project which aimed on recording floral visitors in some common plants in a strategic area for Medellín city in Colombia: and *2)* "Pollination of the cycad *Zamia incognita* A. Lindstr. & Idárraga by *Pharaxonotha* beetles in the Magdalena Medio Valley, Colombia" (*Zamia incognita* dataset), which came from the raw data gathered during a systematic study of the pollination system in a natural population of the endangered and endemic Colombian species of *Zamia* (Valencia-Montoya et al. 2017).

After reviewing the available information for each data set (i.e. plants and insects identities, photographs and field notes for the San Miguel project; and original data about thermogenesis, cones development and insect visitors specimens, for *Zamia incognita* dataset) and exploring the REBIPP template as well as the proposed relational model from Salim et al. (2022), we consider that each dataset could be used to test each model, i.e., dataset 1 was extracted to the REBIPP template and dataset 2 was extracted to the new data model IPT with the biotic interactions publishing model.

3.6.2 Strategy for data standardisation

For the San Miguel dataset, data was extracted to the REBIPP data template using the PPI vocabulary. The original dataset came from two field trips to the "Refugio de Vida Silvestre Alto de San Miguel", and considering that it was placed in the DwC standard based on the occurrence IPT (the one current in production mode within GBIF), some important information for interactions research was underrepresented in some "Remarks" sections. For example, each record contains information from the insect visitor, but relevant plant associated information was originally lost (with plant identification only recorded in the AssociatedTaxa field). Then, considering that each record was linked to a photograph through the record number field, it was possible to rescue this plant's relevant information (e.g. floral phenology). This dataset was thus converted to the Brazilian Plant-Pollinator Interactions Network (REBIPP) data publishing model, which has some DwC fields and some PPI fields. The template was then reviewed in detail for getting used to the available fields and vocabulary; thereafter, records from the standard DwC were just copied and pasted in the new template, and photographs were revisited to get the previously ignored information. Metadata was easily obtained from a work document which has the fieldwork methodology, and the next step will be upload the photographs to a repository, to link them to each dataset record.

The *Zamia incognita* dataset was standardised using the demo version of the integrated publishing toolkit provided by GBIF with the biotic interactions publishing model. Some time was invested in reviewing the new model together with Jose Salim; thereafter, in some meetings the logic scheme of tables' linking IDs was defined (this is fundamental to maintain the data associated). Considering that the cones' temperatures, other measurements, and associated insect visitors came from the same natural population, the main linking ID (EventID) was assigned to each cone from which information was available (e.g., UCES:Maceo:Zi is the ParentEventID for the EvenID UCES:Maceo:Zi:micro009 which belongs to a pollen strobilus, "microestróbilo"in Spanish - male cone). Then all data were placed together in each of the relational tables; cones temperature and

development measurements were placed in the OrganismInteractionAssertion table, while cones visitors data were placed in the MaterialEntity table and their related Assertion fields (the associated information of collected specimens came from digitised specimens from the Terrestrial Arthropods sub-collection, at Colecciones Biológicas de la Universidad CES in Medellín, Colombia - CBUCES-F http://grscicoll.org/institutional-collection/colecciones-biológicas-de-la-universidad-ces-terrestrial-a rthropods). Additional information from gene sequences will be also incorporated in future work.

Finally, considerations for other pollination projects were discussed on the light of the available fields in this relational scheme; for example, some field work also consider scents chemistry, insect visitors or flower structures measurements associated with descriptions of the pollination system; and we concluded that the GBIF proposed new model is flexible enough to hold all these kinds of information.

3.6.3 Time investments

For the San Miguel dataset: we did not need to invest time in cleaning data because the source DwC was already clean, nonetheless most of the time (5 hours in total, 2.5 hours per person) was invested to check each photo, for rescuing some floral traits or phenological stages that were previously lost when the field notes were placed in the Remarks field. We also took some time (3 hours in total, 1.5 hours per person) adjusting the dataset to the proposed controlled vocabulary, because it was made for each record.

For the *Zamia incognita* dataset we invested around 2 hours firstly fitting the insect specimens available information in the REBIPP template; however, considering that we got the original (raw) data associated with the pollination described system we decided to use those data to test the new GBIF relational model. Then we took 2 hours for understanding the relational model exposed in the Salim et al. (2022) paper; thereafter we had a half an hour meeting with José Salim and Debora Drucker to review the relational tables and specific fields of the template, and 2 additional hours to understand the logics of the relational IDs and build them consistently. Once the IDs were filled the specimens based records were pasted in half an hour and temperature as well as development records were standardised and placed in 3 hours; thereafter Organism, OrganismInteraction, and MaterialEntityAssertion data were standardised and complemented with InteractionTypeID links in 1.5 hours. After completing each phase we invested some minutes reviewing the integrity of the data, in sum it was nearly 2 hours (6 hours in total, 2 hours per person). Two additional hours were invested in publishing the dataset through the GBIF ipt and trying to document the metadata. The total time invested for the GBIF new model was 19.5 hours.

It is important to mention that both datasets were already in a digital format (it was not necessary to fulfil spreadsheets again) and that people who prepared the San Miguel dataset were used to working with other DwC archives (for at least two years) and therefore it was not too difficult to understand the template and the needs to fulfil it. The strategy for filling out the relational tables to build the *Zamia incognita* dataset was designed by two persons which have been working with DwC

fields for at least six years, and basic functions in Excel spreadsheets (e.g., concat, transpose, clean spaces and so on) for standardising descriptive information about cones stages or even dates, was already well known for all the team, then they were implemented very quickly.

3.6.4 Lessons learned and recommendations

We think that it is important to take detailed field notes when researchers are recording the interaction; in addition, for insect collections those raw data would enormously enhance the specimens value, therefore an important recommendation for invertebrate collections' managers and curators is to keep scanned copies of the researchers field notebooks (as is usual in vertebrate collections). Also, we recommend that the person who completes the template is aware of the basic characteristics of interactions between organisms. We did not find terms to represent important floral visitors and/or pollinators in wild ecosystems in tropical regions. Although the PPI vocabulary was very useful, in some cases it is not possible to make a match between the original data and the recommended vocabulary. It can be enriched with terms that better represent beetles, moths, flies, hummingbirds and other animals, or terms from other vocabularies can be used (e.g. Hymenoptera Ontology, Uberon, Anatomy of the Insect ontology). In addition, for gimnosperms which do not have flowers and in which biotic pollination is known to be a rule for some groups (e.g., Zamiaceae) it is important to consider additional terms (e.g. micropillar drop, pollen strobilus).

The nature of our first pilot dataset showed us that a great source of insect-visitors and plant interactions could come from iNaturalist and other datasets that can be indexed by GloBI or other platforms; however, the usage of these datasets which come from photographs should be restricted to datasets that have been reviewed and curated by regional botanists and entomologists who surely could enhance enormously the quality of the information (geographically circumscribed) that is extracted from iNaturalist taxonomic identification. For instance, some specific stages of floral phenology can be unambiguously stated from a high-resolution field photograph and the taxonomic resolution of certain insects can be increased by "geographic-fauna" experts.

The proposed biotic interactions publishing model should have the fields definitions and it would be very helpful to have an archive that holds the usage of the relational tables; it is very important to describe the logic framework that should be used to build IDs (i.e., ParenEventID, EventID, AssertionID, and other linking IDs). Even more, the herein tested templates could include recommendations for amateur users: for example, some works can have informative photographs (frequently stored in private email-associated drives) that could be linked to a dataset, whose association could enhance the data quality; nonetheless, robust and open access repositories are needed, and many people could not be familiarised with them. On the other hand, relations ontology was extremely useful.

Finally, the decision of which of the proposed templates a user should use can be addressed through the nature of the dataset; for example, for systematic complete pollination ecology studies,

which too often hold different kinds of measurements and observations, the new GBIF template is recommended.

4. Basis for Guidelines and Recommendations for Publishing Agricultural-related pollinator data

The pilot studies approach allowed for working with a great variety of plant-pollinator data that is produced in this domain and provided the basis for the development of guidelines and recommendations for future initiatives that would like to make their data practices more FAIR. We present some key points learned from this effort below.

4.1 Celebrate Diversity

The pilot projects represent diverse approaches to collect and digitise data (e.g., plant-pollination interactions data, pollinator data) as well as a wide range of supporting institutions (e.g., private companies like HiveTracks, public agencies like USDA, or academic research institutes like the Universidad CES). To embrace this diversity in data and their origins, many approaches to data integration and access rights are needed. For instance, some pilot contributions are openly accessible whereas others, for good reason, are restrictive in their access policies. In some cases, some parts of the pilot are open-access (e.g., metadata is open), whereas the access to the raw data records are restricted (e.g., because they belong to theses in progress). Through our pilot studies, we facilitated discussion on access methods and their benefits. For instance, where pilots with open access policies may benefit from increased visibility, feedback and contributions, pilots with more restrictive policies may be able to better align their data sharing approach with their institutional policies or agreement with their collaborators. So, we recommend to embrace the diversity of pollinator data sources and allocate time to understand the history, capacities and institutional cultures that enabled the collection of valuable pollinator data.

4.2 Embrace Principles of Biodiversity Data Management

Our experience with the pilot projects confirm that knowledge about (meta-)data standards such as DwC and EML help better integrate datasets with different origins. This is why we recommend that professional societies (e.g., funding agencies, journals, and educational institutions) should incorporate principles of biodiversity data management (e.g., standardisation, curation, publication, archiving) in their continued education, curriculums, handbooks, and funding requirements. As currently some journals require, for example, that the field sampling was supported by collecting permits, in the near future they could also require that the data to be published follow the FAIR principles.

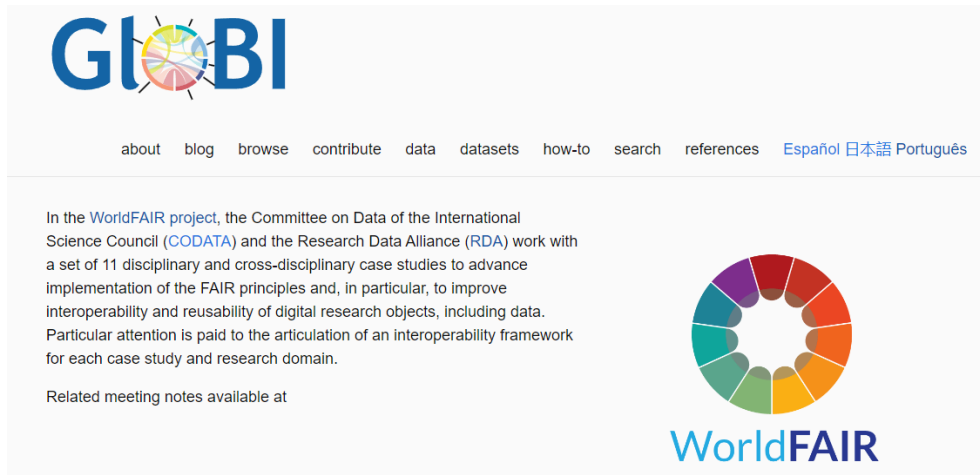## 4.3 Invest in Data Curation, Integration, and Peer-Review Infrastructures

Just like natural history collections actively maintain specimens in their collections, digital data needs constant attention to assure they can be **F**ound, **A**ccessed, **I**ntegrated, and **R**eused. Just like physical specimens can be damaged, misplaced or restored, their digital twins are subject to degradation due to hardware failures, human error, software upgrades, or, more recently, nefarious ransomware attacks. To help keep track of our pilot projects, a dedicated WorldFAIR webpage in the GloBI platform was created at https://www.globalbioticinteractions.org/worldfair/, illustrated in Figure 7. This page provides various perspectives into the pilot projects. These perspectives include a data review paper, which will be described in more detail in Deliverable 10.3, a link to a publicly accessible search index for species interaction data, a reference to versioned controlled pilot metadata, and a pointer to a pilot specific issue tracker (or discussion forum). In addition, the access policies are summarized in metadata, data, and review categories. As mentioned earlier (4.1 Celebrate Diversity), pilots may opt to restrict access to raw data records while allowing for others to inspect their review reports. Finally, pilot contacts are listed to help facilitate direct communication.

Not unlike established practices in natural history collection management (e.g., curation), and scholarly publication processes (e.g., peer-review, editing), the basic tools offered in the WorldFAIR Agricultural Biodiversity Case Study pilot tracking page facilitated peer-review, and data curation of digital pollination records. Anecdotal evidence suggests that these tracking tools help to increase the FAIRness of the pilot data through self-inspection and peer review. Given the dynamic nature of plant-pollination interactions data, we expect that constant curation and periodic review is needed not only to guard the integrity of individual datasets, but also to monitor the connectedness of a vast corpus (or collection) of datasets. In working with biodiversity data in general, and plant-pollination interactions data in specific, connectedness of datasets can be measured quantitatively (e.g., taxonomically, geographically, or temporally overlap) after the first hurdle of data interoperability has been cleared. However, substantial domain expertise is needed to safeguard the integrity and availability of pollination data - just like it takes an expert to perform a taxonomic review of a bee species, domain experts are needed to curate, review, and publish, corpora of versioned pollinator datasets.

To safeguard integrity and access to plant-pollination interactions data, we recommend revisiting the current methods of dataset review and data publication with a specific emphasis on the dynamic nature and interconnectedness of digital data. We encourage academic publishers, research librarians, natural history collection curators, and researchers to work towards a more holistic and continuous approach to compiling, curating, and periodically publishing, citable and trusted digital works such as a corpus of pollination data - the current model of piecewise publication of static pieces of digital data needs attention in order to address Elton's 1927 concern:

"The advantage, and at the same time the difficulty, of ecological work is that it attempts to provide conceptions which can link up into some complete scheme the colossal store of facts about natural history which has accumulated up to date in this rather haphazard manner. [...] Until more

organised information about the subject is available, it is only possible to give a few instances of some of the more clear-cut niches which happen to have been worked out." - Charles Elton, 1927, Animal Ecology. pp 65-66. https://biodiversitylibrary.org/page/7236467

Figure 7a and 7b Screenshots of the GloBI dedicated webpage providing an overview of WorldFAIR associated pilots.

4.4 Track Evidence of Reuse

On being able to find, access and integrate existing data, evidence of data reuse is done by describing, in great detail, *how/where/when* you accessed *what* data and how they connect to the derived work. Only with a systematic approach to documenting reuse, we can track evidence of reuse. Unfortunately, current methods for documenting reuse are limited to including datasets in a reference list, and even if authors go through the effort (anecdotally most don't, as documented in D10.1) to cite all their data in great detail, some academic publishers restrict the number of citations for a given work. Also, current practice of citing works by DOIs may help the counting of references to a particular work, but DOIs are not designed to point to data. So, instead of having detailed coordinates to the content of data, most DOIs point to a web page containing some human readable html content instead of providing a detailed map to specific data. So, to track evidence of reuse, we need to be more precise and systematic in citing data, especially when dealing with complex compilations of datasets such as plant-pollinator records. The example below documents the reuse of a single record from one of our pilot studies. This example describes data reuse in human readable form *and* machine readable form. While a full description of verifiable and secure documentation of reuse goes beyond the scope of this report (see Elliott et al. 2020, 2023; Carvalheiro et al. 2024 for background), our example should provide a glimpse into what is needed to systematically trace evidence of reuse at scale.

4.4.1 Track Evidence of Reuse of a single Plant-Pollinator Record

In 2008, Carvalheiro et al. published the paper based on a plant-pollinator dataset acquired through field observations in England in the period 2004/2005. In their 2008 paper, the original field observations were not included as a supplement. Fortunately, the main author retained their original data in personal collections, and was able to share these files with this Case Study in WorldFAIR.

On line 30 of the file containing the original field observations (Carvalheiro et al. 2024)., a detailed account of an ant (*Lasius alienus*) visiting a flower of a rare plant (*Trinia glauca*) is found. Details included time of recording, place of recording, method used, number of visitations and more. Also, the original data contained a description of the fields used in the table, as illustrated in Figure 8.

Netcode=Carvalheiro England Gully 1, **Study**=Carvalheiro England, **Responsable**=Carvalheiro, **Country**=England, **total.area_sampled_m2**=1480,**METHOD**=Timed observations, **network_type**=All visitors,**Habitat**=calcareous?grassland, **Period**=1, **Site**=Gully, **Plant**=Trinia glauca,**sampling_effort..number.of.flower.units..cm2.definition..observed.**=41, **min_observation**=20, **Notes.on.Minutes.of.observation**=Timed observations, **Total.sampling.effort..fu_x_min.**=820, **Flower.abundance.standardized**=1,38, **units**=number of flower units _cm2_per m2,**FV_Slname**=Lasius alienus, **FV_order**=Hymenoptera, **FV_family_group**=Halictidae, **Visitor_type**=ant, **Visitation.frequency.original**=1, **units.1**=Number of visitors, **Visitation.frequency_stand_fu_min**=0,001219512

Figure 8. Flower visitation data record with column names in **bold** and associated row values separated by commas. Record was extracted from line 30 of the xlsx resource with signature hash://md5/a2b31050b50d9e213ed62873f17c6e8e as described in Carvalheiro et al. 2024. See also https://linker.bio/hash://md5/a2b31050b50d9e213ed62873f17c6e8e.xlsx .

On inspection of the table, records were converted to a consensus REBIPP template described earlier in this report. The REBIPP template resulted from years of collaboration within the network as part of an effort to harmonise plant-pollinator data in Brazil and beyond (Salim, 2023, SURPASS2 project[5]).

The conversion resulted in an associated record in line 468 of a version of the produced interaction data table of the REBIPP template (Carvalheiro et al. 2024). In this example, we were able to match the original record from line 30 and the converted record in line 468 were linked by comparing values like plant name (*Trinia glauca*), animal name (*Lasius alienus*), and the value 41 of column "sampling_effort..number.of.flower.units..cm2.definition..observed." (original) and sampling effort (converted).

plant.recorded.by=Carvalheiro, **plant scientific name**=Trinia glauca, **plant kingdom**=Plantae, **plant rank**=species, **interaction country**=England, **interaction country code**=GB, **interaction habitat**=calcareous grassland, **interaction sampling protocol**=Timed observations, **interaction sampling effort**=41 flowers/cm2, **interaction type**=has flowers visited by, **animal recorded by**=Carvalheiro, **animal scientific name**=Lasius alienus, **animal order**=Hymenoptera, **animal family**=Halictidae, **animal rank**=species

Figure 9 data record derived from record shown in Figure 8 following the REBIPP template as described in line 468 of resource with signature hash://md5/c358eabd4b6921597f1bb3c73e6f5a8c in Carvalheiro et al. 2024. See also https://linker.bio/line:hash://md5/c358eabd4b6921597f1bb3c73e6f5a8c!/L1-L3,L468.

In addition to converting the original flower visitation data into the REBIPP interaction data record template, their associated metadata (e.g., publication date, reference citation, taxonomic range, geographic range) was recorded in a metadata table and subsequently converted in the EML.

Dataset Title=Plant-flower visitor network from Avon Gorge, UK, **Creator name**=Luisa Gigante Carvalheiro, **Organization name**=Universidade Federal de Goiás, **Keywords**=plant-pollinator interactions, flower visitation, **License Name**=Creative Commons Attribution 4.0 International, Abstract=[...], **Geographic Description=**Avon Gorge, Bristol, England, **Begin Date=**2004-05-10, **End Date=**2004-09-27, **Classification System**=[...], **General Taxonomic Coverage**=All flower visitors

---

detected in the study area (Hymenoptera, Diptera, Coleoptera, Heteroptera, Lepidoptera, Thysanoptera), [...]

Figure 10 Comma separated metadata descriptors associated with Carvalheiro et al. 2008 with field names in **bold** with field values following. The metadata fields were truncated to fit into this report, and the full record with signature hash://md5/cf41a46b0c42100413506bf4132a1ac0 can be found via Carvalheiro et al. 2024 and at https://linker.bio/hash://md5/cf41a46b0c42100413506bf4132a1ac0.

Following, EML table definitions were introduced to define a mapping between the REBIPP template and a GloBI supported tabular format for recording species interactions (see https://github.com/globalbioticinteractions/template-dataset).

For instance, the example lines 224-228 and 984-988 of the EML file with signature below shows how the plant scientific name REBIPP value was mapped to the sourceTaxonName GloBI concept.

```
224: <attribute id="sourceTaxonName">

225:    <attributeName>Scientific Name</attributeName>

226:
<attributeDefinition>http://rs.tdwg.org/dwc/terms/scientificName</a
ttributeDefinition>

227:    <storageType>string</storageType>

228 </attribute>

...

984: <attribute id="targetTaxonName">

985:    <attributeName>Scientific Name</attributeName>

986:
<attributeDefinition>http://rs.tdwg.org/dwc/terms/scientificName</a
ttributeDefinition>

987:    <storageType>string</storageType>

988: </attribute>

...
```

Figure 11. lines 224-/228 and line 984-/988 extracted from table definition described in Calvalheiro et al. 2024 eml.xml with signature hash://md5/644e726d2cd6ea9e926e9e2f50e172d8 . The

attribute id "sourceTaxonName" annotates the column with name "Scientific Name" followed by a mapping to attribute id "targetTaxonName" by the columns with the same name from the REBIPP template. See also https://linker.bio/line:hash://md5/644e726d2cd6ea9e926e9e2f50e172d8!/L224-L228,L984-L988 .

After placing the EML file in a github repository tagged to be indexed by GloBI, automated review and indexing processes continuously revisit the mapping configuration and their associated data so that the Carvalheiro data can be reviewed and made searchable through the GloBI services, as illustrated in Figures 12a and 12b..

Figure 12.a Search request and 12.b search results for ants *Lasius alienus* visiting flowers of *Trinia glauca* associated with a version of the Global Biotic Interactions search index as obtained on 2024-01-16 via https://globalbioticinteractions.org. Note that a reference to Carvalheiro et al. dataset if found along with the results from another datasource, The Database of Pollinator Interactions (DoPI, Balfour et al. 2023) making the same claim. Interestingly, when following the

provenance (or origin) of the DoPI interaction claim, the original study of Carvalheiro et al. 2008 appears.

Authors	Title	Methology	Pollinator Species	Plant Species	Interactions	Habitat	Record
L. G. Carvalheiro et al. (2008)	Pollinator networks, alien species and the conservation of rare plants: Trinia glauca as a case study	DO	Lasius alienus	Trinia glauca	1	Calcareous Grassland	38573

Figure 13 Webpage of the Database of Pollinator Interactions (DoPI) detailing the origin of the claim that ant *Lasius alienus* visits flowers of *Trinia glauca* to be Carvalheiro 2008. Because the interaction claim from a single source appears in two different datasets, we have evidence to suggest that Carvalheiro's data is available for reuse.
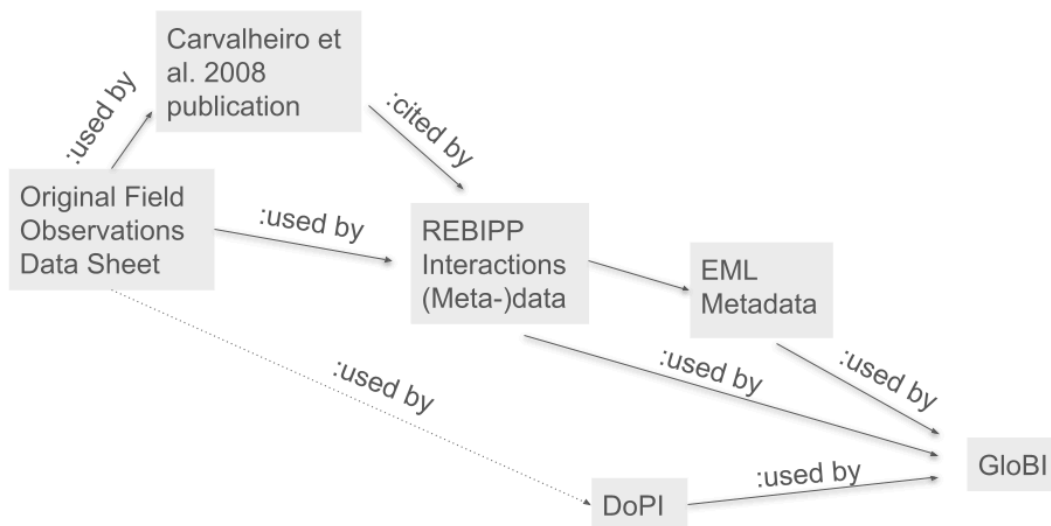

Figure 14. Chain of re-use related to original field observations as referenced in Carvalheiro et al. 2008 as described for a single interaction claim in 2.1.1. Currently, the process of transforming original data into REBIPP interaction (meta-)data and EML metadata is manual and time consuming. Similarly, tracing the provenance (or use) of original field data is a time consuming, manual exercise, except for the automated data processes implemented by GloBI.

Our example in tracing the steps from original data to their reuse suggests that methods are in place to enable reuse of original data. And, our previous example is consistent with experiences reported by our pilots in the following sections: data exchange and reuse patterns rely on manual communication (e.g., email, cloud storage) and transformation processes (e.g., manual or semi-automated conversion workflows). In addition, as we make systematic and appropriate steps towards getting better at documenting how original data sources are reused, we hope to come up

with alternate, and possibly more suitable, ways to facilitate the reuse of valuable plant-pollinator data.

4.5 Cookbook Guidelines and Recommendations for Publishing Plant Pollinator Interactions Data

Lessons learned by pilots provided insights to produce a cookbook with Guidelines and Recommendations for Publishing Agricultural-related pollinator data. The cookbook aims to help different actors standardise plant-pollinator interactions data in accordance with the FAIR principles.

The cookbook was developed using JupyterBook and it is open to anyone to contribute. It introduces the FAIR principles and the WorldFAIR project, and also provides simplified definitions for Persistent Unique Identifiers and the FAIRification process (aka the process of creating FAIR enabled datasets). The FAIRification process chapter presents three different approaches for making a dataset FAIR:

- GloBI: uses GloBI infrastructure and ezeml[6] tool to create FAIR enable datasets.
- REBIPP: uses REBIPP template and information system for creating FAIR enabled datasets.
- GBIF: uses IPT release candidate (RC) 3 for creating FAIR enabled datasets considering the upcoming GBIF "Unified Data Model".

Data authors can select the approach that aligns with their needs and expertise, utilising different tools and systems. However, the ultimate objective is to generate datasets for plant-pollinator interactions that adhere to the FAIR principles.

The guide can be accessed at https://zedomel.github.io/worldfair-agrobio.

5. Collaboration with GBIF - Biotic Interactions Publishing Model

The Case Study on Agricultural Biodiversity contributed to the development of the Biotic Interactions Publishing Model, part of the New Data Model[7], an effort by GBIF, which leads the Case Study on Biodiversity (WP9). A joint webinar[8] entitled "Exploring interactions data on plant pollination" was promoted to disseminate this effort. Further information about the New Data Model in the context of WorldFAIR can be found in Deliverables 9.1, 9.2 and 10.1. In this section, we present the Biotic Interactions Publishing Model and an example of its implementation with one of the pilot datasets.

The biotic interactions publishing model defines each interaction as a DwC Event between two DwC Organisms (Figure 15). In that sense, the publishing model considers interactions at organism level,

[6] https://ezeml.edirepository.org/
[7] https://www.gbif.org/new-data-model
[8] https://www.gbif.org/composition/6yJQMg2cvnnzb88xwt62m7/exploring-interactions-data-on-plant-pollination

opposed to the species level interactions. Additional characteristics of interacting organisms or the interactions themselves can be included using *Assertions* of the New Model (an entity similar to DwC MeasurementOrFact). DwC MaterialEntity can be used to include details about the specimens of interacting organisms, and GeneticSequence may be used to include genetic details about the organisms.
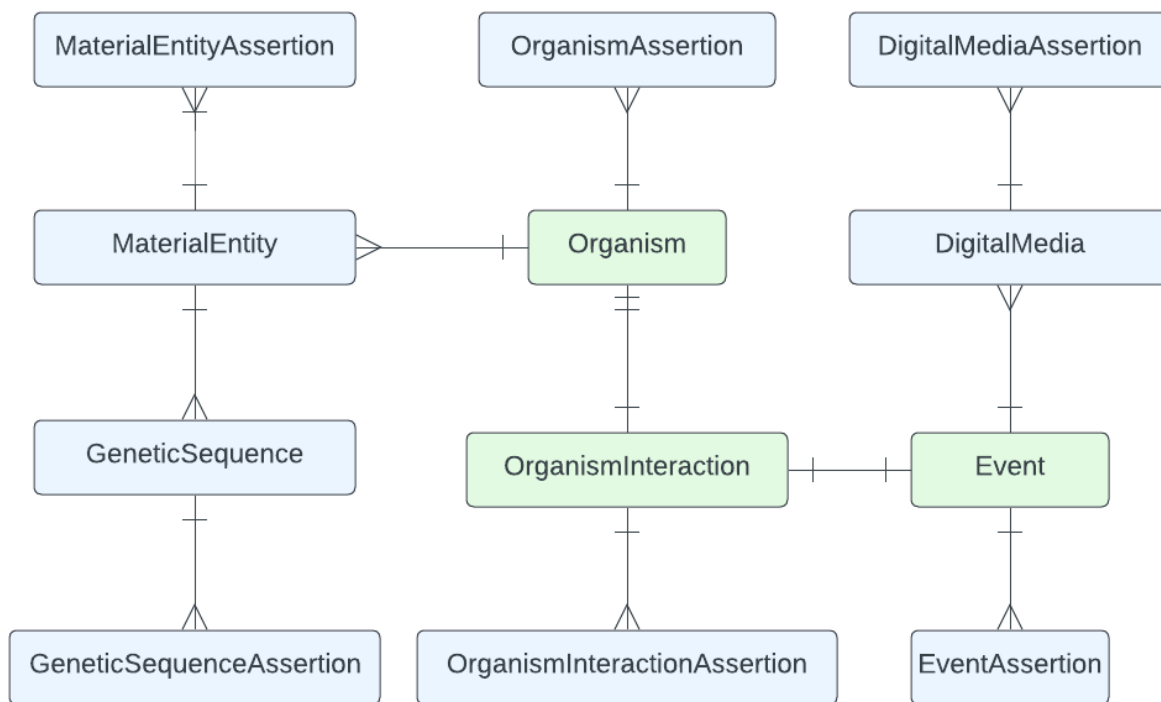

Figure 15: Schema of biotic interactions publishing model for the GBIF New Data Model and IPT. In green are the mandatory data and in blue the optional data.

We selected two pilot datasets to test the publishing model and the new version of IPT, which incorporates several new use cases, including biotic interactions. The *carvalheiro2023* dataset was used to populate the tables of the publishing model and then it was published using an IPT installation created for this project (https://worldfair-ipt.gbif-uat.org/). The other dataset came from original (raw) data from the description of the pollination system in a wild population of *Zamia incognita* (pollination ecology description was already published: Valencia-Montoya et al. 2017).

We present a screenshot of the Demo GBIF IPT in Figure 16, but that no resources are currently available, indicating that the GBIF IPT is not yet being used, or the resources are not visible.

The first step was to populate the publishing model template with data from carvalheiro2023 data set. Only the tables Event, OrganismInteraction, Organism and OrganismInteractionAssertion were needed for this dataset

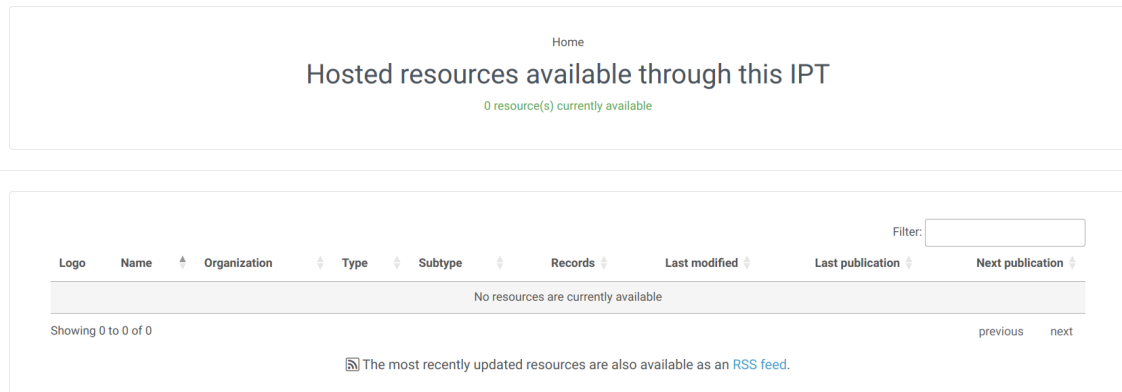([https://docs.google.com/spreadsheets/d/1EPJXXCdc4OFDU0QW2zzlkT7Bv05z4jBhKJ3K3GClTEY/edi](https://docs.google.com/spreadsheets/d/1EPJXXCdc4OFDU0QW2zzlkT7Bv05z4jBhKJ3K3GClTEY/edit#gid=0)[t#gid=0](https://docs.google.com/spreadsheets/d/1EPJXXCdc4OFDU0QW2zzlkT7Bv05z4jBhKJ3K3GClTEY/edit#gid=0)).

Figure 16: Screenshot of GBIF IPT https://worldfair-ipt.gbif-uat.org/ as accessed on 12 January 2024.

In the second dataset, the tables OrganismInteraction, OrganismInteractionAssertion, Event, Organism, MaterialEntity and MaterialEntityAssertion were used from the *Zamia* dataset [https://docs.google.com/spreadsheets/d/1gk-JHgRm209NmB9ln24V0clXKAq3hVpO/edit?usp=shari](https://docs.google.com/spreadsheets/d/1gk-JHgRm209NmB9ln24V0clXKAq3hVpO/edit?usp=sharing&ouid=115614938304847609114&rtpof=true&sd=true)[ng&ouid=115614938304847609114&rtpof=true&sd=true](https://docs.google.com/spreadsheets/d/1gk-JHgRm209NmB9ln24V0clXKAq3hVpO/edit?usp=sharing&ouid=115614938304847609114&rtpof=true&sd=true)

Each table was exported to CSV files, since IPT works with this file format for data importing.

In IPT, the first step is to create a new data package and provide a name for the dataset and the respective type as Darwin Core Interaction Package - Interaction DP (Figure 17).
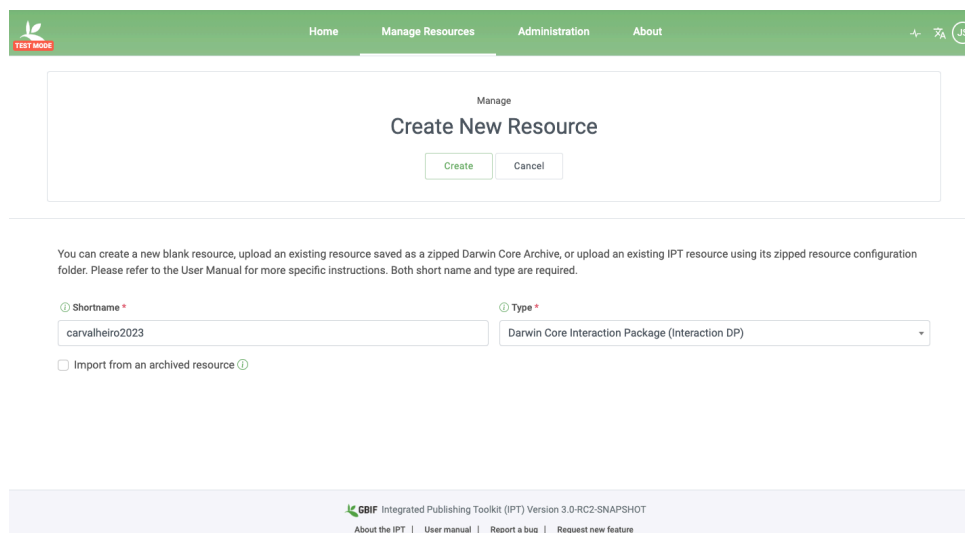

Figure 17: Create new resource in IPT for biotic interaction dataset.

Next step involves uploading the CSV files generated from the publishing model template: event.csv, organisminteraction.csv, organism.csv and organisminteractionassertion.csv (Figure 18). After files have been uploaded, the data mappings between columns in the original CSV files to DwC terms need to be set. Since the publishing model already uses DwC terms as column labels the data mappings are significantly simplified, and it is only necessary to map the Interaction DP file to the source files (Figure 19). After all source files have been mapped to a respective table in the publishing model, the metadata must be filled before publishing the resource (Figure 20) and the visibility set to public.
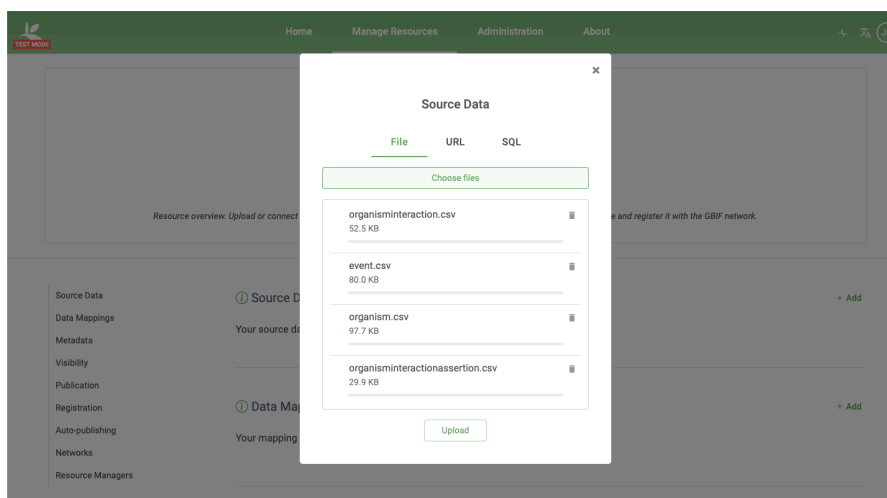

Figure 18: Setting source data for the Interaction DP in IPT.
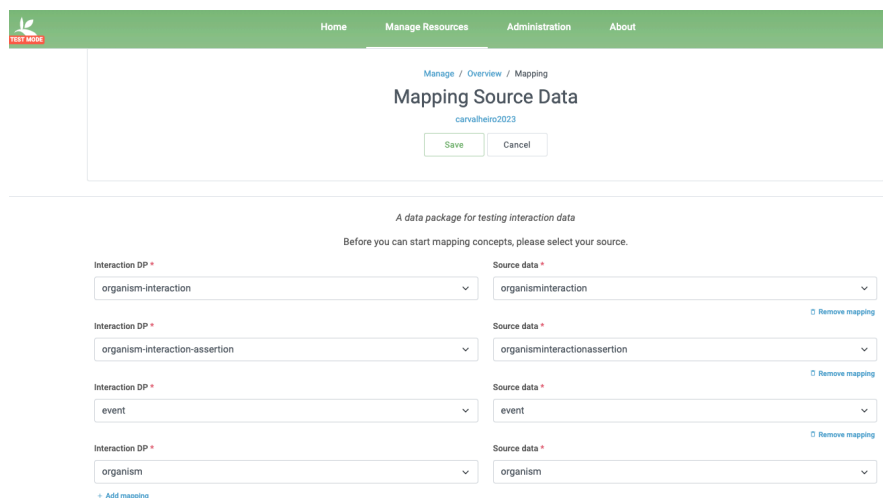

Figure 19: Data mapping for the organism-interaction table.

In Figure 19, the source file *organisminteraction.csv* is mapped to the table organism-interaction of the publishing model (Interaction DP).

Figure 20: Basic metadata in IPT.

The demo version of the IPT was tested successfully and, once in production by GBIF, will allow for data providers to easily publish plant-pollinator interactions data, as well as other biotic interactions events. Despite the IPT publishing model's complexity, particularly when compared to other solutions, the Integrated Publishing Toolkit (IPT) enables a more extensive and comprehensive documentation of biotic interactions. This includes diverse data types such as genetic sequences, assertions, digital media, and physical material, surpassing what can be effectively represented

using flat files. This effort is a result of a fruitful collaboration between Work Packages 9 and 10 in WorldFAIR and their will to contribute to Biodiversity data interoperability and sharing.

6. Cost of adoption estimates

Each of the pilots used a [collaboratively developed template](#) to estimate how many hours were invested into the various activities within each pilot. These activities ranged from executive buy-in on the front end all the way through writing the final report for the pilot. These estimates, in turn, would allow for a good projection regarding how much time would be required to do a similar adoption of the data standards, such as the Darwin Core standard, in another context. In addition, each organisation provided the number of team members working on the pilot, a quick overview of the organisation(s) behind the pilot, and, where applicable, the type and scale of the data used in the pilot. This information would then allow for the time investments by each pilot to be better contextualised within the available resources and goals of the organisation(s) behind each pilot.

Notably, the activities that received the greatest investment of time summed across all the pilots included:

- Learning the data standards
- Cleaning and standardising the data retrieved
- Writing a relevant tutorial as a part of the pilot
- Writing WorldFAIR Reports

However, many of the pilots varied greatly in how they distributed their time investment reflecting the unique goals and, sometimes, processes for each pilot. This can be visualised in Figure 21 where the time invested into each activity is summed across all pilots and then segmented by pilot:

Hours Invested In Each Activity Segmented by Pilot

Figure 21: Pilot-reported estimates for how many hours each pilot spent on each activity.

In Figure 22 we can also visualise approximately what percent of the total hours spent were used on each activity (assessed as a percent breakdown within each pilot averaged across pilots to avoid being affected by large differences between pilots in total hours spent):

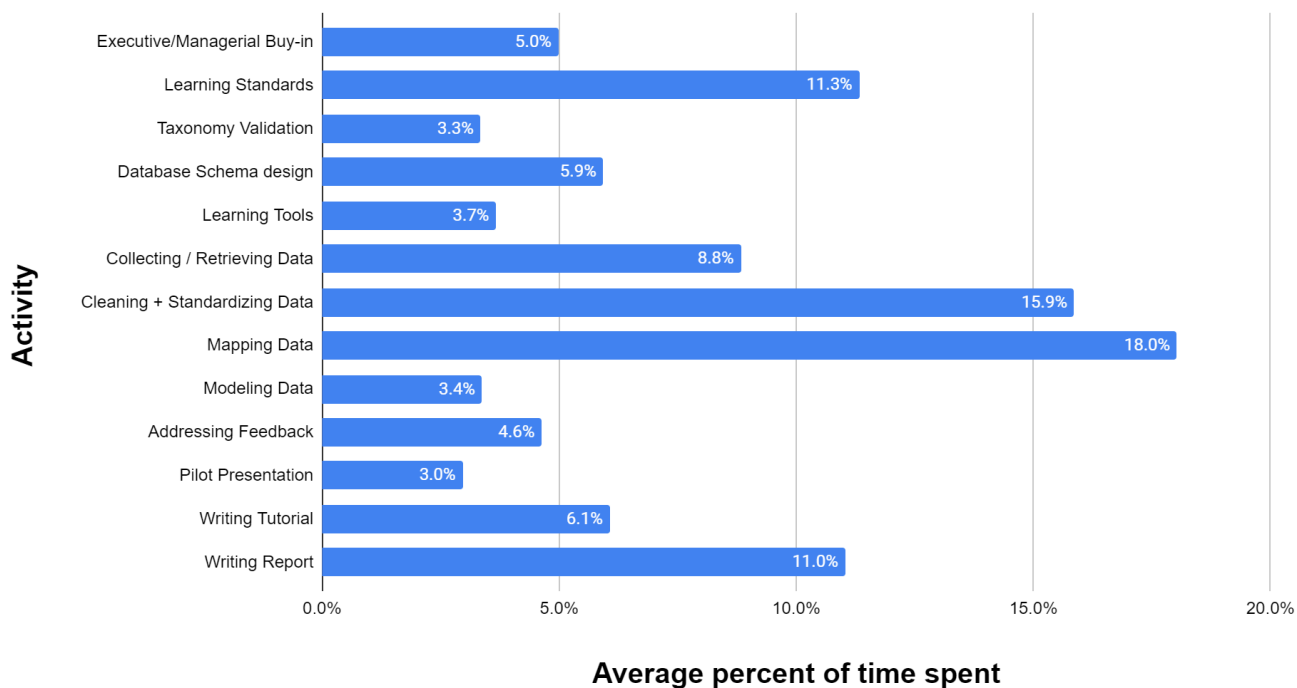## Average Time Spent On Each Activity Across All Pilots

Figure 22: Average percent of time spent on each activity based on pilot-reported estimates

7. Semantic Interoperability and CDIF

The Cross Domain Interoperability Framework (CDIF) provides guidelines for achieving cross-domain interoperability.

> The idea is to identify capabilities and component services, along with the information payload - with appropriate models and interfaces--to act as a lingua franca within and between domain and infrastructure boundaries. The framework includes modules for various aspects of interoperability and reuse, including discovery (this module), data integration, semantic harmonisation, data access, and the supporting technology (CDIF Working Group 2023 p.3).

The discovery module (CDIF Working Group 2023) presents recommendations for making resources discoverable on the web. Regarding the recommendations for metadata content, the metadata model we adopted in this project conforms to some of the recommendations. The metadata schema selected for creating records, which were subsequently imported by GloBI, is the EML schema. An example of a metadata record can be checked at https://linker.bio/hash://md5/cf41a46b0c42100413506bf4132a1ac0. This metadata record has

been mapped using terms from various domain-agnostic standards, specifically Schema.org, FOAF (Friend of a Friend) ontology, DCAT (Data Catalog Vocabulary), and Dublin Core. Additionally, to preserve essential semantic details, terms from the Darwin Core standard were also incorporated, despite it not being domain-agnostic. An example of this enriched metadata record is available at https://github.com/globalbioticinteractions/carvalheiro2023/issues/1#issuecomment-1855661190. In this metadata record, we can see the following recommended metadata attributes recommended by CDIF (CDIF Working Group 2023):

- Resource identifier and resource type: <https://docs.google.com/spreadsheets/d/1cJ0qX9ppqHoSyqFykwYJef-DFOzoutthBXjwKRY81T8/edit#gid=359918449> rdf:type sdo:Dataset ;
- Title: dcterms:title "Plant-flower visitor network from Avon Gorge, UK" ;
- Rights: dcterms:rights "Creative Commons Attribution 4.0 International" - additional details about access rights were provided on the GloBI page where the datasets were indexed (see Fig. 7). sdo:license <https://creativecommons.org/licenses/by/4.0/> was used to provide the license URI;
- Description: sdo:description "This dataset gathers information on interactions between plants [...];
- Originators: Specified as follows:

sdo:creator [

  rdf:type foaf:Person ;

  foaf:name "Luisa Gigante Carvalheiro" ;

  foaf:mbox <mailto:lgcarvalheiro@gmail.com> ;

  foaf:based_near [

    rdf:type foaf:Location ;

    foaf:city "Goiania" ;

    foaf:country "Brazil"^^<http://www.w3.org/2001/XMLSchema#string> ;

  ] ;

  foaf:affiliation [

    rdf:type foaf:Organization ;

    foaf:name "Universidade Federal de Goiás" ;

   ] ;

 ] ;

sdo:publisher [

  rdf:type foaf:Organization ;

  foaf:name "Universidade Federal de Goiás" ;

- Temporal Coverage:

dcat:startDate "2004-05-10"^^<http://www.w3.org/2001/XMLSchema#date> ;

dcat:endDate "2004-09-27"^^<http://www.w3.org/2001/XMLSchema#date> ;

- Geographic Extent: dcterms:spatial "Avon Gorge, Bristol, England"^^<http://www.w3.org/2001/XMLSchema#string> ;

Additional metadata related to provenance and dataset distribution were integrated during the data processing phase within GloBI. However, it's important to note that these elements are not showcased in the provided example. This example does not include all the recommendations for metadata content from CDIF Working Group (2023). Future discussions are planned to explore the description of plant-pollinator interactions data, specifically focusing on enhancing cross-domain interoperability using CDIF.

8. Recommendations

Recommendation 1 - Embrace Principles of Biodiversity Data Management

Recommendation type:

- ○ *Technical (data)*
- ○ *Technical (metadata)*
- ○ *Organisational*

Stakeholders: Researchers, Data Stewards, Project Managers, Repository Managers, Funding Agencies, Journal Editors, Educational Institutions

We recommend professional societies (e.g., funding agencies, journals, and educational institutions) to incorporate principles of biodiversity data management (e.g., standardisation, curation, publication, archiving) in their continued education, curriculums, handbooks, and funding requirements. Namely, adoption of (meta)data standards such as Darwin Core (DwC) - https://dwc.tdwg.org/, the Ecological Metadata Language (EML) - https://eml.ecoinformatics.org/, and the Plant-Pollinator Interaction vocabulary (PPI) - https://ppi.rebipp.org.br/list/, may help better integrate datasets with different origins. These standards are utilised by many institutions worldwide, demonstrating their broad applicability and acceptance in the scientific community. There exists active communities that engages in the ongoing development and refinement of these standards. This community is characterised by their transparency and participatory approach to curation, ensuring that the standards continually evolves to meet the dynamic needs of their users. The adoption of these standards, as we demonstrated in this project, enable interoperability with major biodiversity data platforms such as the Global Biodiversity Information Facility (GBIF), and the Global Biotic Interactions (GloBI). In addition to these metadata standards, the Relation Ontology (RO) - http://purl.obolibrary.org/obo/ro.owl played a pivotal role in the standardisation of plant-pollinator interaction data. This ontology provided specific terms, such as ro:has_flowers_visited_by and ro:visits_flowers_of, which were instrumental in defining the directional relationships between animal and plant species within the datasets. The use of these terms ensured a more accurate and consistent representation of interspecies interactions.

Recommendation 2 - Invest in Data Curation, Integration, and Peer-Review Infrastructures

Recommendation type:

- ○ *Policy*
- ○ *Organisational*
- ○ *Technical (data)*

Stakeholders: Researchers, Data Stewards, Project Managers, Repository Managers, Funding Agencies, Journal Editors, Educational Institutions

We recommend investments in data infrastructures that facilitate data curation, integration and peer-review. To safeguard integrity and access to plant-pollination interactions data, we recommend revisiting the current methods of dataset review and data publication with a specific emphasis on the dynamic nature and interconnectedness of digital data. We encourage academic publishers, research librarians, natural history collection curators, and researchers to work towards a more holistic and continuous approach to compiling, curating, and periodically publishing, citable and trusted digital works such as a corpus of pollination data. Given the dynamic nature of pollination data, we expect that constant curation and periodic review is needed not only to guard the integrity of individual datasets, but also to monitor the connectedness of a vast corpus (or collection) of datasets. In working with biodiversity data in general, and pollination data in specific, connectedness of datasets can be measured quantitatively (e.g., taxonomically, geographically, or temporally overlap) after the first hurdle of data interoperability has been cleared. However, substantial domain expertise is needed to safeguard the integrity and availability of pollination data - just like it takes an expert to review a taxonomic review of a bee species, domain experts are needed to curate, review, and publish, corpora of versioned pollinator datasets.

9. Conclusions

This report presents results from the pilot phase of the Agricultural Biodiversity Case Study, describing efforts from six different initiatives to adopt standards recommendations from the discovery phase. The pilots allowed for dealing with concrete and diverse examples and for the generation of reusable materials tailored to this domain, as well as for obtaining better estimates for the overall costs of adoption for future projects.

We have successfully promoted the adoption of standards and increased the interoperability of plant-pollinator interactions data, resulting in a process that allows for tracing the provenance of the data, as well as facilitating the reuse of datasets which are extremely important to understand this essential ecosystem service and how it is changing due to human impact.

Our approach was flexible enough to deal with a variety of approaches to plant-pollinator interactions data which we believe are comprehensive of the universe of data in this domain that is out there, encompassing private companies like HiveTracks, public agencies like USDA, or academic research institutes like the Universidad CES. In the academic realm, different methodologies were applied, covering experiments and field observations in several regions and croplands as well as preserved ecosystems.

Our effort shows there are several possible paths for FAIRification, according to the institutional needs, and we demonstrated that different approaches can sum up to promote data interoperability and data availability for reuse, which is the ultimate goal of this initiative. As such, we have successfully achieved the goal of ensuring FAIR data for understanding plant-pollinator interactions at biologically-relevant scales for crops, with broad participation from initiatives in Europe, South America, Africa, North America and elsewhere. We also established concrete guidelines on FAIR data best practices customised for pollination data, metadata and other digital objects and promoted scalable adoption of these standards and FAIR data best practices by multiple initiatives. We believe this effort can help similar initiatives to adopt interoperability standards for this domain and contribute to our understanding of how plant-pollinator interactions sustain life on Earth.

10. Appendix

Guidelines and Recommendations for Publishing Agricultural-related pollinator data - https://zedomel.github.io/worldfair-agrobio/

Tutorial to standardise a plant-pollinator dataset with OpenRefine, to be shared in REBIPP - https://drive.google.com/drive/folders/17k6NNNyHvebv6TGH-KYb8NF_q3S_cP11?usp=drive_link

11. Bibliography

BeeXML – Exchanging Data about Bees and Beekeeping. (n.d.). BeeXML. Retrieved January 12, 2024, from https://beexml.org/

Benson, T., & Grieve, G. (2021). Why interoperability is hard. In T. Benson & G. Grieve, Principles of Health Interoperability (pp. 21–40). Springer International Publishing. https://doi.org/10.1007/978-3-030-56883-2_2

Carvalheiro, LG; Barbosa, E.R.M. & Memmott, J. 2008. Pollinator networks, alien species and the conservation of rare plants: Trinia glauca as a case study. Journal of Applied Ecology, 45,1419-1427. DOI: https://doi.org/10.1111/j.1365-2664.2008.01518.x .

Carvalheiro, L. G., Soares, F., Salim, J. A., Poelen, J.H., & Drucker, D. (2024). Provenance of WorldFAIR pilot data from: VisitationData_Luisa_Carvalheiro. https://docs.google.com/spreadsheets/d/1cJ0qX9ppqHoSyqFykwYJef-DFOzoutthBXjwKRY81 T8 hash://md5/048875415b7cc9fb27f1189a8a946ff5 hash://sha256/dec6efdd95fd64d5c38480e0db0dfa329c94e8e0fc0736f0769cafb470fd13ce (0.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10530109

CDIF Working Group, Richard, S., Gregory, A., Hodson, S., Fils, D., Kanjala, C., Bell, D., Winstanley, P., Edwards, M., Heus, P., Brickley, D., Rizzolo, F., Maxwell, L., Luis, G., Buttigieg, P. L., & Le Franc, Y. (2023). Cross Domain Interoperability Framework (CDIF): Discovery Module (v01 draft for public consultation) (Version 01). Zenodo. https://doi.org/10.5281/zenodo.10252564

Nick Balfour, Maria Clara Castellanos, Chris Johnson, Dave Goulson, Andrew Philippides. 2023. The Database of Pollinator Interactions (DoPI). Accessed at https://www.sussex.ac.uk/lifesci/ebe/dopi/ on 2023-12-01. Accessed via <https://github.com/globalbioticinteractions/dopi/archive/9574bd55b2689b1b393abfad882 9c2c6b1567ed0.zip> at 2023-12-23T00:09:57.847Z.

iNaturalist. (n.d.). iNaturalist. Retrieved January 12, 2024, from https://www.inaturalist.org/

IPBES. (2016). The Assessment Report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on Pollinators, Pollination and Food Production. Potts, S.G., Imperatriz-Fonseca, V.L. and Ngo, H.T. (eds). Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Bonn, Germany.

Monasterolo, M., Poggio, S.L., Medan, D., Devoto, M., 2020. Wider road verges sustain higher plant species richness and pollinator abundance in intensively managed agroecosystems. Agric. Ecosyst. Environ. 302, 107084. https://doi.org/10.1016/j.agee.2020.107084

Pl@ntNet API for developers. (n.d.). Pl@ntNet API. Retrieved January 12, 2024, from https://my.plantnet.org/

Jorrit H. Poelen, James D. Simons and Chris J. Mungall. (2014). Global Biotic Interactions: An open infrastructure to share and analyse species-interaction datasets. Ecological Informatics. https://doi.org/10.1016/j.ecoinf.2014.08.005.

Salim J.A. et al. (2022). Data standardisation of plant–pollinator interactions, GigaScience, Volume 11, giac043, https://doi.org/10.1093/gigascience/giac043

Salim, JA. Unifying biotic interactions data: terminology, data analysis, standardization, and proposal of a data schema for plant-pollinator interactions. 2023. [doi:10.11606/T.3.2023.tde-03012024-110904]. São Paulo : Escola Politécnica, Universidade de São Paulo. Tese de Doutorado em Sistemas Digitais.

Tavella, J., Windsor, F.M., Rother, D.C., Evans, D.M., Guimarães, P.R., Palacios, T.P., Lois, M., Devoto, M., 2022. Using motifs in ecological networks to identify the role of plants in crop margins for multiple agriculture functions. Agric. Ecosyst. Environ. 331, 107912. https://doi.org/10.1016/j.agee.2022.107912

Turkmayali, A. (2023, September 14). Semantic interoperability: A common language for data sharing. International Data Spaces. https://internationaldataspaces.org/semantic-interoperability-a-common-language-for-data-sharing/

Valencia-Montoya, W. A., Tuberquia, D., Guzmán, P. A. & J. Cardona-Duque. (2017). Pollination of the cycad Zamia incognita A. Lindstr. & Idárraga by Pharaxonotha beetles in the Magdalena Medio Valley, Colombia: a mutualism dependent on a specific pollinator and its significance for conservation. Arthropod-plant Interactions 11(5): 717–729. https://doi.org/10.1007/s11829-017-9511-y

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18