Training workshop on standards and prototype data sharing tools for the consortium + external users (M27)

STATUS: FINAL

Project acronym:	EU BON
Project name:	EU BON: Building the European Biodiversity Observation Network
Call:	ENV.2012.6.2-2
Grant agreement:	308454
Project Duration:	01/12/2012 – 31/05/2017 (54 months)
Co-ordinator:	MfN, Museum für Naturkunde - Leibniz Institute for Evolution and Biodiversity Science, Germany
Partners:	UTARTU, University of Tartu, Natural History Museum , Estonia
	UEF, University of Eastern Finland, Digitisation Centre, Finland
	GBIF, Global Biodiversity Information Facility, Denmark
	UniLeeds, University of Leeds, School of Biology, UK
	UFZ, Helmholtz Centre for Environmental Research, Germany
	CSIC, The Spanish National Research Council, Doñana Biological Station, Spain
	UCAM, University of Cambridge, Centre for Science and Policy, UK
	CNRS-IMBE, Mediterranean Institute of marine and terrestrial Biodiversity and Ecology, France
	Pensoft, Pensoft Publishers Ltd, Bulgaria
	SGN, Senckenberg Gesellschaft für Naturforschung, Germany
	VIZZUALITY, Vizzuality S.L., Spain
	FIN, FishBase Information and Research Group, Inc., Philippines
	HCMR, Hellenic Centre for Marine Research, Greece
	NHM, The Natural History Museum, London
	BGBM, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany
	UCPH, University of Copenhagen: Natural History Museum of Denmark, Denmark
	RMCA, Royal Museum of Central Africa, Belgium
	PLAZI, Plazi GmbH, Switzerland
	GlueCAD, GlueCAD Ltd. – Engineering IT, Israel
	IEEP, Institute for European Environmental Policy, UK
	INPA, National Institute of Amazonian Research, Brazil
	NRM, Swedish Museum of Natural History, Sweden
	IBSAS, Slovak Academy of Sciences, Institute of Botany, Slovakia
	EBCC-CTFC, Forest Technology Centre of Catalonia, Spain
	NBIC, Norwegian Biodiversity Information Centre, Norway
	FEM, Fondazione Edmund Mach, Italy
	TerraData, TerraData environmetrics, Monterotondo Marittimo, Italy
	EURAC, European Academy of Bozen/Bolzano, Italy
	WCMC, UNEP World Conservation Monitoring Centre, UK

EU BON

EU BON: Building the European Biodiversity Observation Network
Project no. 308454

Large scale collaborative project

MS282
Training workshop on standards and prototype data sharing tools for the consortium + external users

Milestone number	MS282
Milestone name	Training workshop on standards and prototype data sharing tools for the consortium + external users
WP no.	WP2
Lead Beneficiary (full name and Acronym)	RMCA
Nature	Workshop held
Delivery date from Annex I (proj. month)	2015-02-28 (M27)
Delivered	yes
Actual forecast delivery date	2015-04-17 (M29)
Comments	Annex 1-3 are included

| Name of the Authors	Name of the Partner	Logo of the Partner
Larissa Smirnova	MRAC	
Patricia Mergen	MRAC	
Isa Van de Velde	MRAC	
Israel Pe'er	GlueCAD	
Hannu Saarenmaa	UEF	

In case the report consists of the delivery of materials (guidelines, manuscripts, etc)

Delivery name	Delivery name	From Partner	To Partner

Summary of the Milestone

The Milestone MS282 summarises the results of the **Second EU BON Training Event** and its evaluation based on the participants' feedback. The Second EU BON Training Event was held March 19, 2015, back to back with the joint CETAF-EU BON informatics workshops[1]. The event was organised by the University of Finland (UEF)[2] and Digitarium[3] (WP2 leader), in collaboration with CETAF ISTC and other EU BON work packages. It aimed to continue introducing participants to the information architecture, data standards and data sharing tools developed or used by the EU BON consortium and the accent was made on the hands-on session with GBIF new Integrated Publishing Toolkit (IPT) which was adapted by EU BON to enable the flow of sample-based data in support of EBVs (Essential Biodiversity Variables).

The training was attended by 28 people from the EU BON consortium and beyond (from 13 countries).
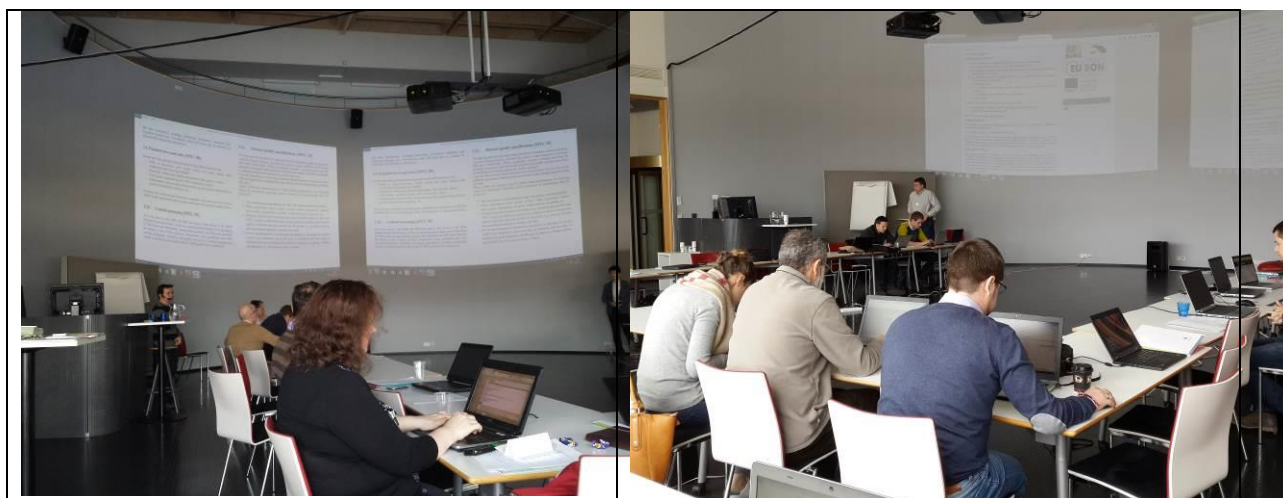
Introduction

This training event on data sharing tools was held to achieve the objectives set by the project:

- *Develop data integration and interoperability between the various networks, and with new generation of data sharing tools enhance linking between observational data, ecosystem monitoring data, and remote sensing data*
- *Develop new web service interfaces for data holdings using state-of-the-art standards and protocols*
- *Ensure global coordination of development efforts through an international data interoperability task force and adoption of the results through helpdesk and a comprehensive training program*

This activity is also aligned with key actions of CBD towards the 2020 targets[4]:

*"Strengthening and promoting the further mobilization of and access to data by, for example, encouraging the use of common informatics **standards** and protocols, promoting a culture of **data sharing**, investing in digitization of natural history collections and promoting citizen scientists' contributions to the body of biodiversity observations".*

---

[1] http://digitarium.fi/en/content/eu-bon-and-cetaf-joint-informatics-workshop
[2] http://www.uef.fi/
[3] http://digitarium.fi/en
[4] www.cbd.int/gbo4

To support the data mobilization from biodiversity observation networks the EU BON project is adapting the existing tools for handling and publishing metadata, occurrence data and ecological data from GBIF, namely the IPT.

Today huge volume of biodiversity data is being generated worldwide by thousands of environmental, ecological and natural resource observations (e.g. different monitoring programs and volunteer-based sightings). Outputs of systematic monitoring schemes, also called Sample-based data are characterized with features such as:

- quantitative data, calibrated measures;
- applying standard sampling protocols;
- relay on standards scientific naming and IDs;
- repeatable sampling, comparable;

and thus can be used in modelling and analysis to detect changes and trends in populations, aim at supporting the EBVs (see presentation of Éamonn Ó Tuama[5]).

Achievements and current status

The training was dedicated to the hands-on session on data publishing using the new EU BON IPT. This tool has been developed further from the existing GBIF IPT (v. 2.2), and is configured to work with a data model for sample-based data (including a new Event core and measurement extensions) and compatible with the DwC-A model. Additionally, the introductory courses were given on GEOSS/GEO BON connection, on the current state of EU BON software architecture that will support the EU BON Biodiversity Portal, and on the Darwin Core sample data model.

The practical part of the training included the demonstration of the new IPT and some examples of sample-based data published by EU BON partners. It was followed by practical exercise where participants could publish data (test dataset or own dataset) using the test version of the tool. The program and presentations are available online[6] and in **Annex 1.**

The training was attended by 28 people from the EU BON consortium and beyond (from 13 countries) (**Annex 2**). 14 people have given their feedback via an anonymous online survey (see **Annex 3**). The reactions of participants after the training were very positive, people have appreciated the possibility to try out the publishing process in real (even while yet in test mode).

The training was rated as good (36%) to very good (43%), the objectives of the training were clearly defined and completed. Most of the participants found that the training experience will be useful for their work (93%) and half of the respondents plan to use it after the training. 75% of participants were beginners, 25% have used the IPT before. Most of them were familiar with the Darwin Core standard (50% advanced, 25% - intermediate and 25% – beginners). The new functionalities of IPT are widely appreciated (see Annex 2 for detailed assessment).

Challenges and further/future developments

Although it was a good starting training with the new IPT, there is still much to improve in the future. Based on the discussion held during the training session and on the feedback of the participants, the next points should be taken into account:

- Create a clearly structured workflow for the process of data publishing, including an assistant guide regarding the data preparation steps.
- Consider to offer a video for illustrating these steps.
- Printed manual is also needed.

---

[5] http://eubon.cybertaxonomy.africamuseum.be/sites/default/files/DwC-sample-data-Joensuu-final.pdf
[6] http://eubon.cybertaxonomy.africamuseum.be/node/1648

- For the demonstration of publishing sample-based data – involve more actively the test sites, ask them to provide samples of their data and to try out the EU BON IPT, so that issues can be revealed, solved and described.
- Use-cases are important to demonstrate how to publish various kinds of data (including different sampling methods, measurements, size etc.).
- The hand-on session can be extended and different types of datasets could be used (simple versus advanced).
- If people bring own data, request a copy beforehand, so the trainers can study this and discuss special issues on the data (if any).
- For the hands-on training several trainers (mentors) should be present to help with solving issues and answering particular questions.
- Basic technical background of IPT is necessary (requirements, user account, rights, registration of organization, export-import, DOI, etc.).
- Darwin Core is a very complex standard, would be good to have a manual for DwC mapping, including new terms for sample-based data (also illustrated by use-cases).
- Consider a separate dedicated training session on data mapping.

Other important aspects could also be highlighted or discussed, such as:
- Other data sharing tools, like DEIMS (for test sites) and PlutoF (for citizen science).
- Illustration on how published sample-based data can be used in EBVs.
- How to attract data provider to extend existing dataset with additional (environmental) information.
- How to mobilize East European data providers.

For the upcoming trainings the next action points were planned:
- Keep track of all workshops and trainings organized / held within and outside the EU BON consortium.
- Have this information placed on the Data mobilization helpdesk.
- Consider different possibilities for the next training (DoW: Training workshop on data standards and sharing tools for external users worldwide, due month 33): JRC ISPRA, GBIF Madagascar, TDWG Kenia. There is manifestation of interest from Crete and Tartu.

Annex 1: Topics covered during the training session

Introduction to GEOSS, GEO BON, EU BON[7] (Hannu Saarenmaa):

The presentation covers origins, organisation, and current plans of the Group on Earth Observation (GEO), the goals and information architecture of the Global Earth Observation System of Systems (GEOSS). GEOSS Portal, registry system, data sharing principles, available data, and brokering mechanisms is explained. Participation in the GEOSS Architecture Implementation pilot process is discussed. The GEO Biodiversity Observation System (GEO BON) and its aims for information management was also covered.

Information architecture of EU BON[8] (Antonio Garcia):

The presentation covers the current update or the software architecture that will support the EU BON Biodiversity Portal, focusing on the brokering alternatives and data/metadata sharing standards (EML, OGC-CSW, SOAP and REST interfaces..). GI-cat is introduced as a brokering tool to integrate new data sources. This session also includes a demonstration of the first implementation prototype of the EU BON Biodiversity Portal.

Data standards, Darwin Core and extensions for sample-based quantitative data[9] (Éamonn Ó Tuama):

The Darwin Core vocabulary, extended with a small number of additional terms, can be used in a Darwin Core Archive to encode information from sample-based data sets, i.e., data sets associated with environmental, ecological, and natural resource investigations. Such data are usually quantitative, calibrated, and follow certain protocols so that changes and trends in populations can be detected. This session introduces the new terms, the star schema model underlying Darwin Core archives consisting of a core table linked to one or more extension tables, and the associated enhancements to the GBIF Integrated Publishing Toolkit (IPT) to support publishing of sample-based data.

Demonstration of GBIF/EU BON IPT for monitoring networks[10], [11], [12](Larissa Smirnova and Franck Theeten):

The Integrated Publishing Toolkit (IPT) is a software tool developed by GBIF, aiming to facilitate the sharing and publishing of biodiversity data on the Internet using the GBIF network. It uses the Darwin Core standard to map species occurrence datasets and checklists, and can also handle data from natural sciences collections or observations. Since additional, sample-based terms are added to the DwC vocabulary, the IPT tool can be used by various monitoring networks collecting mainly quantitative data (environmental, ecological, and natural resource investigations). The practical part of the training will present the IPT tool, and explain how to publish your dataset using the IPT tool, via

---

[7] GEOBON_GEOSS_infra-v5.pdf

[8] 2nd_Training_EU_BON_Architecture_v03.pdf

[9] DwC-sample-data-Joensuu-March2015.pdf

[10] EU BON_Presentation_IPT tool.pdf

[11] eubon_vocabularies_only_Theeten.pdf

[12] eubon_ipt_only_Theeten.pdf

a practical example using occurrence data. Extensions of the IPT will be presented and issues will be discussed after the training.

Practical exercise with sample dataset (Larissa Smirnova and Franck Theeten)

Presentation of Israel Peer on example of sample-based data[13]

Practical exercise with own data (all trainees)

---

[13] EventBaseExample-gluecad (1).pdf

Annex 2: List of participants (* -EU BON partner, ** -Associate partner)

	Name	Country	Institute	Job title
1	Aaike De Wever*	Belgium	Royal Belgian Institute of Natural Sciences	Science officer
2	Andreas Allspach*	Germany	Senckenberg Research Institute Frankfurt (SGN)	Collections Manager
3	Ana Casino**	Belgium	CETAF	General Secretary
4	Anton Güntsch*	Germany	Freie Universität Berlin	Head, Biodiversity Informatics Research Group
5	Cedric Chaveriat	France	French Foundation for Biodiversity Research	Developer
6	Christian Langer	Germany	GEO BON	GIS Programmer
7	Aurelie Delavaud**	France	French Foundation for Research on Biodiversity	ECOSCOPE executive manager
8	Éamonn Ó Tuama*	Denmark	GBIF	Senior Programme Officer for Interoperability
9	Elspeth Haston	UK	Royal Botanic Garden Edinburgh	Deputy Herbarium Curator
10	Florian Wetzel*	Germany	Museum für Naturkunde	Researcher
11	Florian Wolf	Germany	iDIV	GIS Assistant
12	Francisco Antonio García Camacho*	Spain	CSIC	Software Engineer
13	Franck Theeten*	Belgium	RMCA	Database manager
14	Gregor Hagedorn*	Germany	Museum für Naturkunde	Head of Digital World
15	Hanna Koivula	Finland	Finnish Museum of Natural History - Luomus	ICT specialist
16	Hannu Saarenmaa*	Finland	UEF, Digitarium	Director
17	Henry Engledow	Belgium	Botanic Garden Meise	Data Base Manager
18	Israel Peer*	Israel	GlueCAD	General Manager
19	Jiri Frank	Czech Republic	National Museum	Head of department of digital asset management and new media
20	Larissa Smirnova*	Belgium	RMCA	Science officer
21	Marian van der Meij**	Netherlands	Naturalis	Information manager
22	Marie-Elise Lecoq	France	GBIF France - MNHN	IT developer
23	Matúš Kempa*	Slovakia	Institute of Botany, Slovak Academy of Sciences	
24	Patricia Mergen *	Belgium	RMCA	Liaison Officer
25	Miguel Fernandez	Germany	GEO BON	Postdoc
26	Sophie Pamerlon	France	GBIF France - MNHN	Biodiversity Data Engineer
27	Teodor Georgiev*	Bulgaria	Pensoft Publishers	Technical Director
28	Veljo Runnel*	Estonia	University of Tartu Natural History Museum	Citizen science specialist

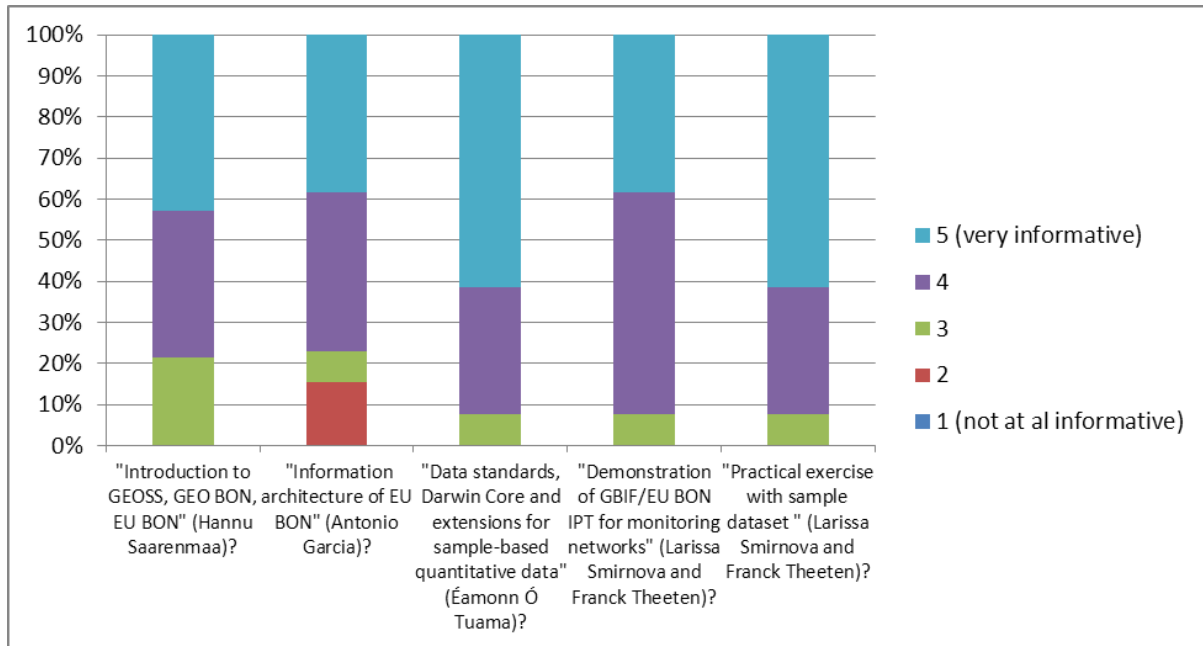## **Annex 3:** Feedback summary from Online-Survey

1. How do you rate the overall training:

5 (very good)	4	3	2	1 (very bad)	Total	Weighted Average
42.86%	35.71%	21.43%	0.00%	0.00%		
6	5	3	0	0	14	4.21
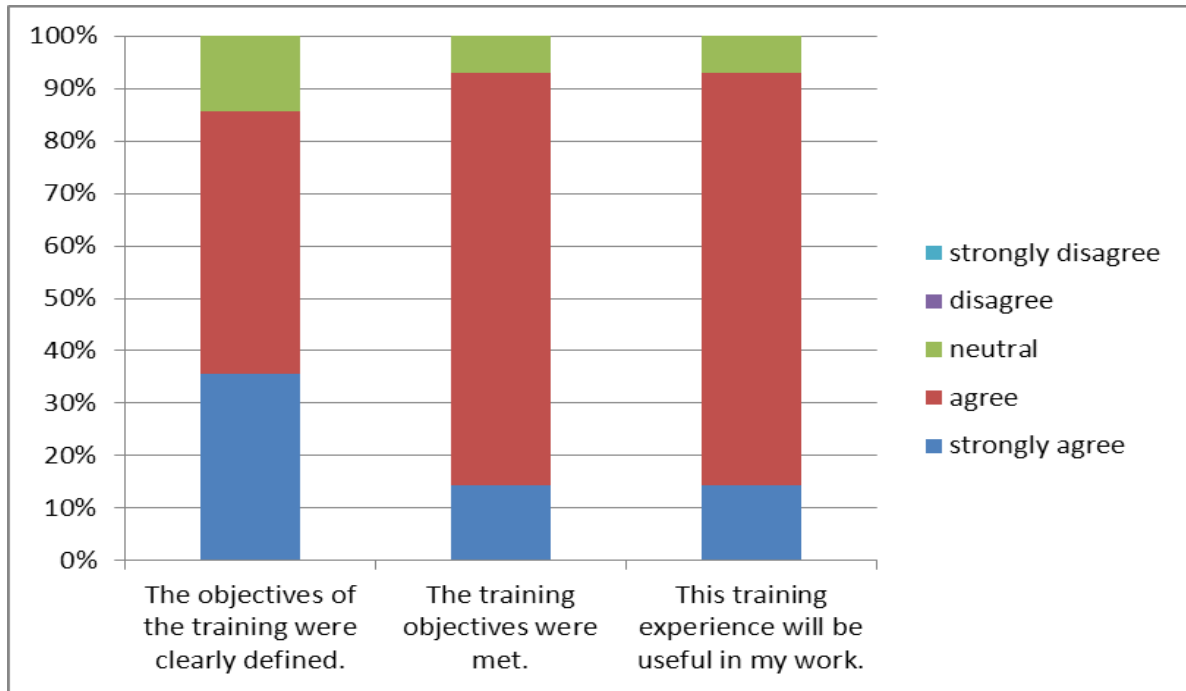
2. Evaluation per module:

How informative was the session…

Do you have suggestions for improvement of this session (Hannu)?
Ok, as general background, but maybe not really required as part of a training more focused on IPT and DwC.

Do you have suggestions for improvement of this session (Antonio)?
Sharing ideas on various issues, more about internals
Not easy to follow. I am not sure whether this has to do with the level of English or the content - I found it difficult to concentrate. It was also not clear why certain things were being done.
This presentation was very focused on the system design but failed to deliver this in an accessible way, and did not discuss the consequences of specific design choices for the users and data providers.
Actually the presentation was information packed, but a lot of this information seemed only relevant to those people deeply involved in the system design. Unfortunately this resulted in the fact that more strategic considerations in terms of integrated data/information sources and the way this data/info is served to the user were not sufficiently dealt with. This presentation was also not well integrated with the rest of the training.

Do you have suggestions for improvement of this session (Eamonn)?
Good presentation. For training purposes I think the presentation could be complemented with some hands-on examples.

Do you have suggestions for improvement of this session (Larissa+Franck)?
A workflow starting from hints on how to prepare the source files and a complete (not full) session of the process.
Follow and planned talk and don't deviate too often - almost got derailed at one point.
Fine, but duo presentation could be improved a bit.

Do you have suggestions for improvement of this session (practical)?
A workflow starting from hints on how to prepare the source files and a complete (not full) session of the process.
Maybe it would be useful to have a range of different training datasets. For the persons bringing their own data, it would be useful to request a copy beforehand, so the trainers can study this and discuss special issues with the data (if any).

What was particularly helpful about the training?
Learn how to use the new version of the IPT
The examples using the EU BON Portals and the GI-CAT. The explanation on IPT's vocabulary. The taxonomic & registry
hands on exercise
Learning how to use the last version of the IPT tool.
To do it yourself, go through the whole process in the practical exercise
The trainers circulated within the group assisting us when we got lost.
Getting to know the specifics of data mapping in GBIF IPT was very good.
The training session on IPT tools . The presentations of the EU BON portal and repository
Having the test accounts so we could do it 'for real'
The discussions and practical examples on the IPT and Darwin Core.

What can be improved in the content of the training? In the training style and organization?
Involve more people from the testing sites
More in-depth on technical issues regarding the architecture
More background for beginners
More working groups & synthesis
We should encourage people from WP5 to attend.
A digital or paper manual for the participants for the practical exercise. More discussion and information about the roadmap of EUBON and other standards
Starting the training by giving a good example of what you could use the tool for. This would then be followed by a clear structure of the major steps to be taken. The example will then be done following the steps outlined. It is important to return regularly to the structure to indicate the position in the process, this helps the learning process immensely. Too much deviation or going off in tangents (even if important) may lead to confusion or people getting lost. Only once the whole process has been explained is it interesting to build or elaborate on the structure given.
More "forced" feedback from participants would have been helpful, discussion groups with task to formulate feedback on special topics etc. In general asking feedback to specific, predefined questions or problems will help.
Written protocols would be useful to refer to afterwards.
As mentioned above: wherever possible, I would suggest to have more hands-on topics and less traditional presentations.

What other subjects for training are you interested in?
Data sharing tools. Workflows and analysis
EBV classes
Data standards
EML, ISO... metadata standards and the way to complete them by fill in form or through API to import / export metadata from others IS.
Data sharing tools and techniques, data analysis.
Limits of the ipt server, is there a maximum file size for instance. If so, how to deal with that.   Definition of fields
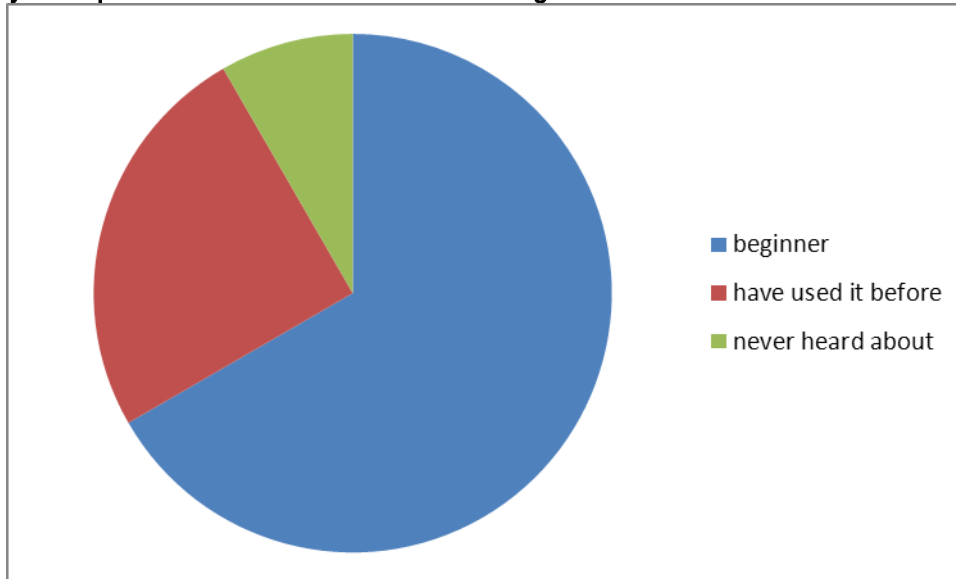Data exporting, importing, data standard converting etc
Data cleaning
I think it would be useful to go more in depth into Darwin Core itself and tackle more examples with (tricky) Darwin Core mapping.

3. IPT-related questions:

What was your experience in IPT use before the training?

- beginner
- have used it before
- never heard about

What do you think about the new functionality of EU BON IPT (core and extensions for sample-based data)?

- important and easy to use
- important, not easy to use
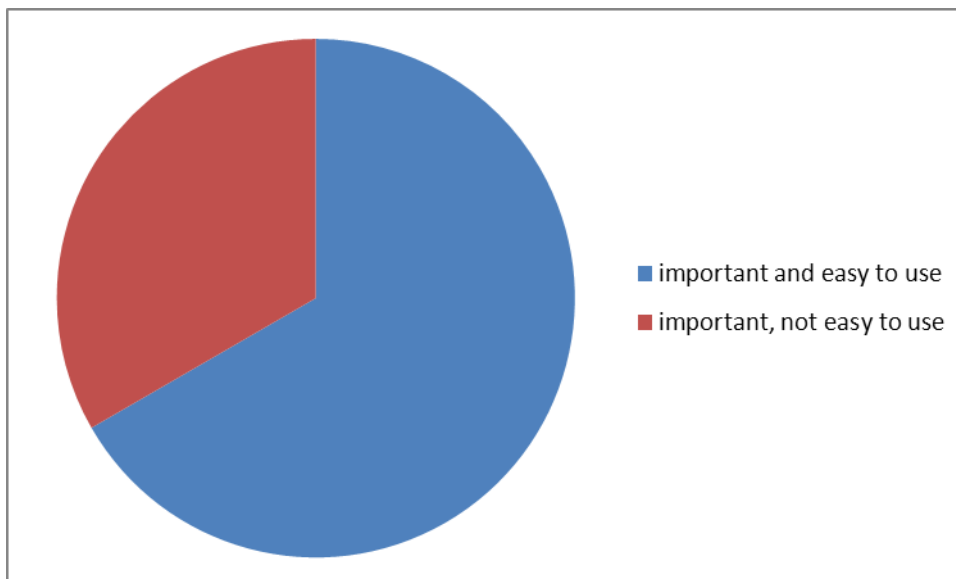
Comments:
It not about the use, it is about the definition of terms used in IPT : some pop up boxes with the definition could improve the use.
It was not so clear to me how the core and extensions interact
Type of core should be renamed from "Other" to "Event"

How easy was it to follow the training presentation?

5 (Easy)	4	3	2	1 (difficult)	Total	Weighted Average
41.67%	50.00%	8.33%	0.00%	0.00%		
5	6	1	0	0	12	4.33

What kind of data would you like to publish with the IPT?
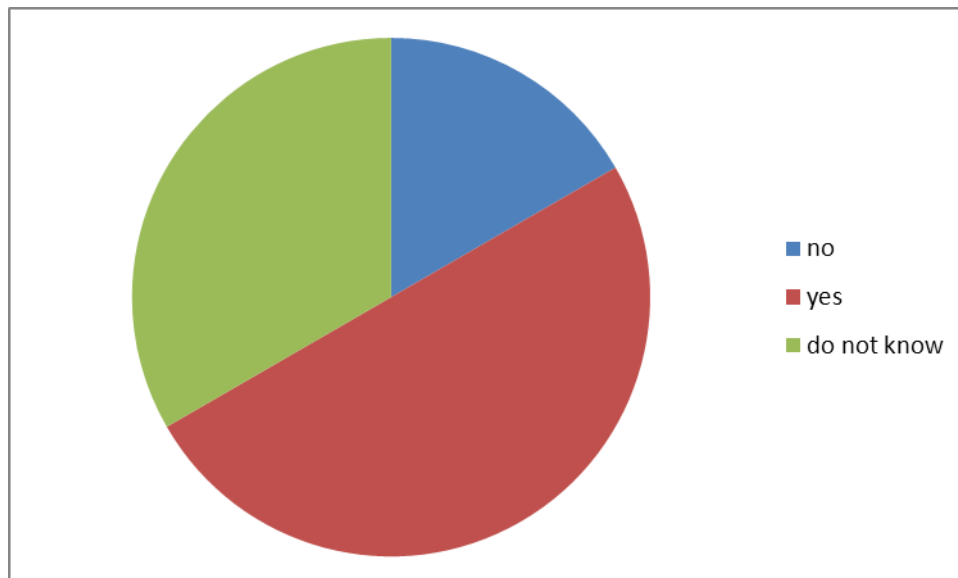Observations and monitoring data
Occurrences
Occurrence data
The Israeli butterflies : 1 Sporadic observations   2. Systematic monitoring scheme
Zoological, geological, paleontological en botanical data
Observation data

Are you planning to use the EU BON IPT in the future?

Comments:
IPT prepares data in DwCA format for GBIF but unfortunately my country does not participate in this infrastructure so I do not expect to use this tool in the future
I prefer a system that automatically updates changes when the "mother" data set is changed. I am not a fan of uploading static data sets. I fear we are creating a lot of duplicate data points.
We won't use it because we haven't got scientific data, but our partners should need it and we can help them to use it
I am not the person responsible for publishing our data to GBIF but I will be able to discuss the IPT with this person and I will let other people know in other institutes who do not currently submit their data to GBIF. I believe it will be extremely useful for them.

4. DwC related questions

What is your knowledge/experience of DarwinCore?

- advanced
- intermediate
- beginner
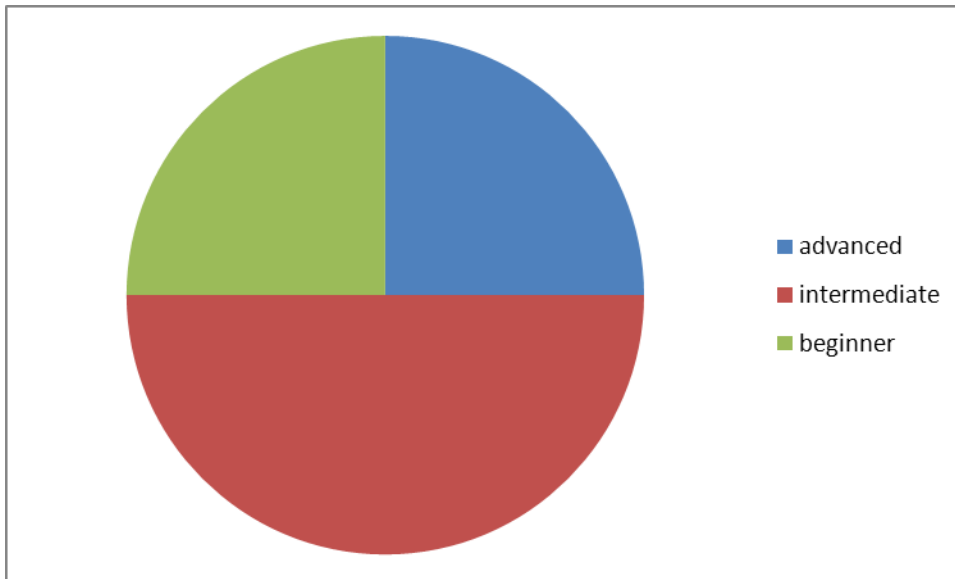
How well does DarwinCore meet your needs?

- Very well
- Rather well
- Not so well
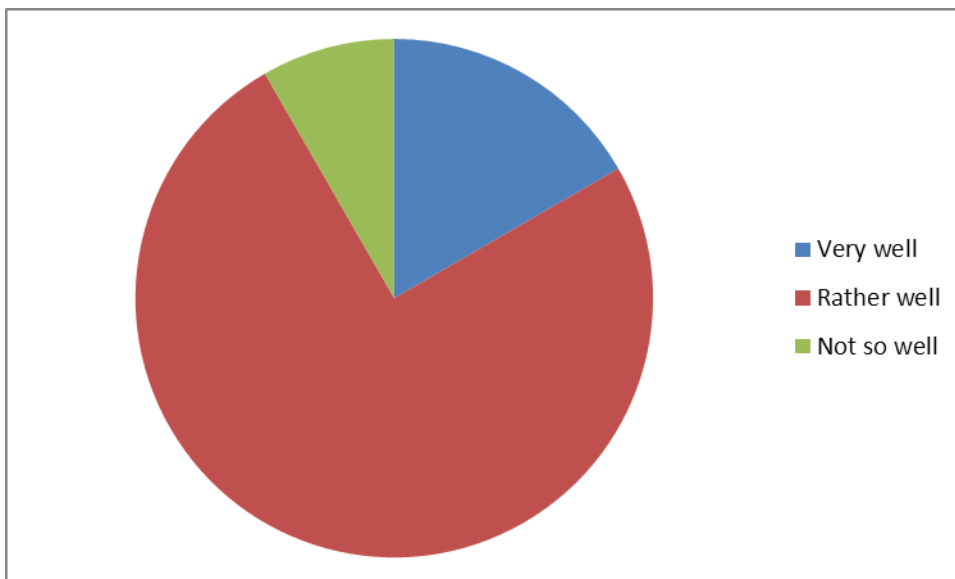
Comments:

It ok when working on the species level of biodiversity. But not at all when working on ecosystem structures and functions or genetic resources. Those two levels should not be forget when talking about biodiversity data.
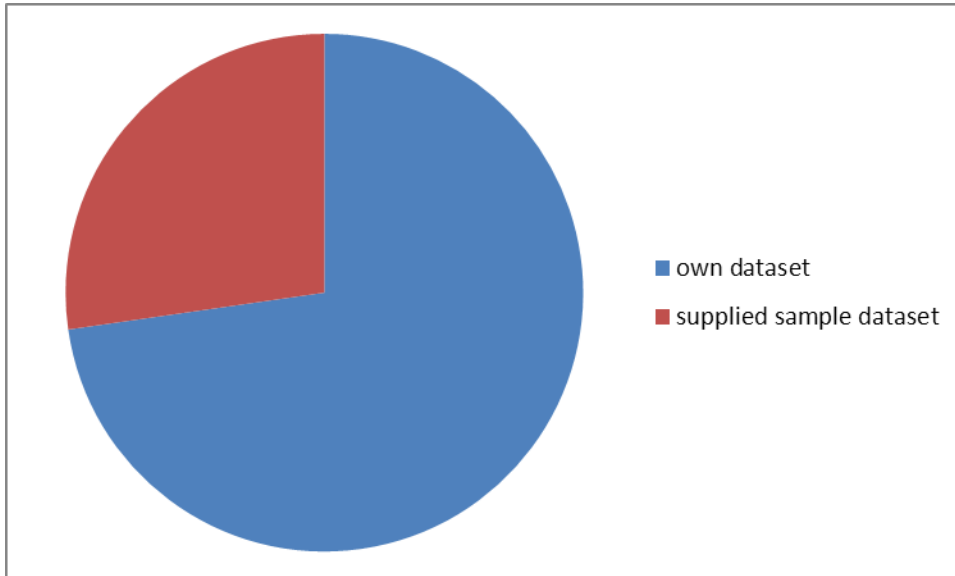
ABCD is more complete.

What terms are missing in current DwC vocabulary?
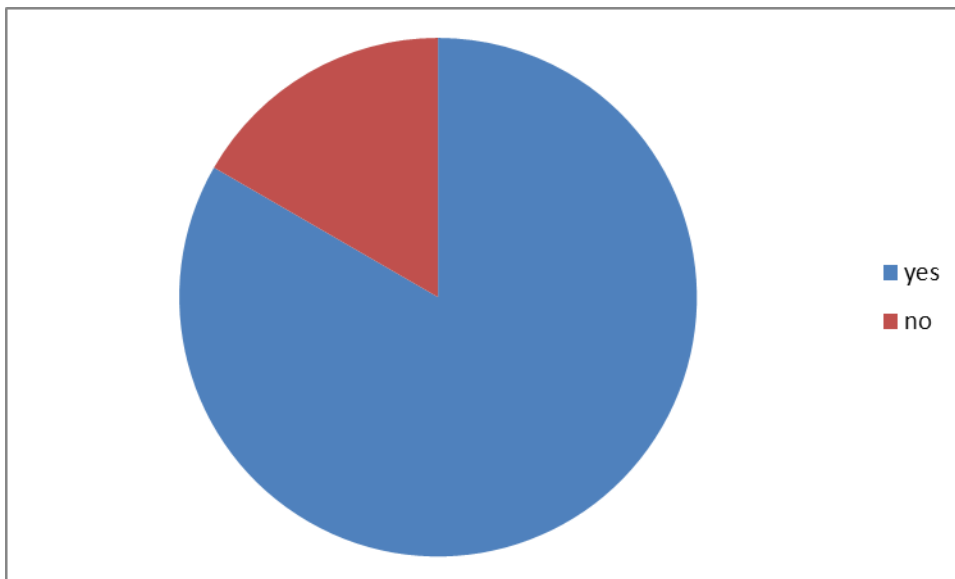
Sampling protocols

Container for mineral en rock determinations

5. Questions about used datasets

What data did you use for the practical exercise?

- own dataset
- supplied sample dataset

Was the supplied sample dataset suitable for the illustration of the training topic?

- yes
- no

Comments:
It may be interesting to do a separate session (before this training) on data and the DARWIN Core Standard. All the explanation of the various fields in the session distracts from the training.
Probably data was too complex?
The training is supposed to learn you something. Different datasets going from easy to more complex ones could help to get "absolute beginners" started.