**Milestone MS281**

Version: 1
Date: 29/04/2014
Author: MRAC
Document reference: Milestone_MS281

# Training workshop on data architecture and standards for up to 25 people from the consortium (M14)

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<br>UEF, University of Eastern Finland, Digitisation Centre, Finland<br>GBIF, Global Biodiversity Information Facility, Denmark<br>UniLeeds, University of Leeds, School of Biology, UK<br>UFZ, Helmholtz Centre for Environmental Research, Germany<br>CSIC, The Spanish National Research Council, Doñana Biological Station, Spain<br>UCAM, University of Cambridge, Centre for Science and Policy, UK<br>CNRS-IMBE, Mediterranean Institute of marine and terrestrial Biodiversity and Ecology, France<br>Pensoft, Pensoft Publishers Ltd, Bulgaria<br>SGN, Senckenberg Gesellschaft für Naturforschung, Germany<br>VIZZUALITY, Vizzuality S.L., Spain<br>FIN, FishBase Information and Research Group, Inc., Philippines<br>HCMR, Hellenic Centre for Marine Research, Greece<br>NHM, The Natural History Museum, London<br>BGBM, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany<br>UCPH, University of Copenhagen: Natural History Museum of Denmark, Denmark<br>RMCA, Royal Museum of Central Africa, Belgium<br>PLAZI, Plazi GmbH, Switzerland<br>GlueCAD, GlueCAD Ltd. – Engineering IT, Israel<br>IEEP, Institute for European Environmental Policy, UK<br>INPA, National Institute of Amazonian Research, Brazil<br>NRM, Swedish Museum of Natural History, Sweden<br>IBSAS, Slovak Academy of Sciences, Institute of Botany, Slovakia<br>EBCC-CTFC, Forest Technology Centre of Catalonia, Spain<br>NBIC, Norwegian Biodiversity Information Centre, Norway<br>FEM, Fondazione Edmund Mach, Italy<br>TerraData, TerraData environmetrics, Monterotondo Marittimo, Italy<br>EURAC, European Academy of Bozen/Bolzano, Italy<br>WCMC, UNEP World Conservation Monitoring Centre, UK |

This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No 308454.

# EU BON

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

Large scale collaborative project

# MS281
# Training workshop on data architecture and standards for up to 25 people from the consortium

| | |
|---|---|
| **Milestone number** | MS281 |
| **Milestone name** | Training workshop on data architecture and standards for up to 25 people from the consortium |
| **WP no.** | WP2 |
| **Lead Beneficiary (full name and Acronym)** | MRAC |
| **Nature** | Written report |
| **Delivery date from Annex I (proj. month)** | 2014-01-31 |
| **Delivered** | yes |
| **Actual forecast delivery date** | 2014-05-05 |
| **Comments** | |

| Name of the Authors | Name of the Partner | Logo of the Partner |
|---|---|---|
| Larissa Smirnova | MRAC | |
| Kim Jacobsen | MRAC | |
| Patricia Mergen | MRAC | |
| | | |
| | | |

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 MS281 reports back the results of the **First EU BON Training Event** and summarises lessons learned as recommendations for the future trainings. The First EU BON Training Event was held on April 3rd 2014, following the General Meeting of EU BON in Crete. It aimed to provide a basic understanding and shared vocabulary on Information architecture and Data standards used by EU BON partners. Six experts from different networks have given an introduction course to up to 25 attendees from the EU BON consortium.

# Introduction

To achieve the objectives set by the project:

- to establish an information architecture compatible with the global GEO BON, INSPIRE, other European projects
- to develop data integration and interoperability between the various networks,

it was decided not to build its information architectures from scratch but to adopt existing relevant infrastructures and approaches from established international initiatives (i.e., GBIF, LTER, GEOSS, GEO BON, LifeWatch, INSPIRE and DataOne) that would fit the EU BON goals. Therefore the basic introduction to the existing projects and networks, overview of relevant infrastructures, demonstration of use of the informatics tools and services developed by different initiatives was identified as a priority for the first training event.

# Achievements and current status

This first training event was aimed at EU BON members and pursued to encourage a basic understanding and shared vocabulary on Information architecture and Data standards used by EU BON. Several topics were covered (http://eubon.cybertaxonomy.africamuseum.be/node/5) (see **Annex 1**):

1. Integrating biodiversity networks through Software Oriented Architecture
2. Data standards: publishing sample-based data
3. Information architecture - GEOSS perspective
4. Data sharing and repositories in GBIF
5. Data flow and modelling in virtual laboratory
6. Data sharing and repositories in DataONE

The training was attended by 15-25 participants (number depends on the course given) from 13 countries and from almost all WPs. The courses were given by 6 trainers - experts from EU BON and from related Networks of Biodiversity Information. Each course had a time slot of one hour. The workshop was organized by Kim Jacobsen and Larissa Smirnova (both MRAC), but chaired by Hannu Saarenmaa (UEF).

Participants and trainers were asked to fill in the online feedback survey in order to evaluate the training and to get a better idea about future needs. We have received 16 answers from the attendees (~50%) and 6 from the trainers (100%). The training is evaluated rather high (see questions/answers in the **Annex 2**) and the feedback is very positive. Participants found the training very useful as it gave a very good overview of existing infrastructures in general and EU BON goals in particular. They also appreciated the practical examples of data sharing possibilities and access and found the overview of different technologies and initiatives inspiring.

# Challenges and further/future developments

In the survey we have asked what other subjects for training could be interesting in the future. Many participants have mentioned the data mining, data analysis (comparability) and data flow as of great importance. Interoperability of data from different domains and architectural aspects are also points of interest. More hands-on sessions and practical involvement will be highly appreciated. Due to differences in IT level of participants one of the regular comments is to organise less technical workshop. But also from the trainer's side there is a suggestion to give two different workshops: one for the "already technical" people and the other for beginners.

In case of organising more practical sessions: it will require more preparation work, setup and possible specific logistic arrangements, for example on network bandwidth. It was also noted that such training should be better organised at the beginning of the meeting, rather than at the end. A common request for future trainings was to provide more training material (also to put online) than the PowerPoint presentations, for example video's with practical demos, manuals, etc.

The next training workshop (MS282) is expected to be in February 2015 and will be open for the consortium and external users.

# **Annex 1:** Titles and Abstracts of the training workshop

### 1. Integrating biodiversity networks through Software Oriented Architecture - Antonio Garcia (CSIC)

**Presentation:** EU BON_EAI_SOA-presentation.pdf

In this training session we will introduce the concepts of Software Oriented Architecture, Business Process Modelling and Enterprise Application Integration, analyzing their relevance with EU BON architectural design (D2.1) and presenting different ways to accomplish the integration of biodiversity networks and other data sources. We will present a demonstration of a working system that will integrate several data sources through and Enterprise Services Bus.

### 2. Data standards: publishing sample-based data using the GBIF Integrated Publishing Toolkit - Eamonn O Tuama (GBIF)

**Presentation:** EUBON-crete-sample-data.pdf

The GBIF Integrated Publishing Toolkit (IPT) is the recommended application for publishing data to the GBIF network. To date, it has been used to publish three types of data: taxon occurrences, checklists and data set level metadata. In this session, we will explore its adaptation for publishing sample based data. First, we will review the essential attributes of sample data that need to be captured. Then we will introduce the Darwin Core Archive data format, explain the constraints imposed by its star – schema, relational data model, and address the requirement for additional terminology in the Darwin Core vocabulary to describe the attributes of sample data together with controlled value vocabularies for some attributes. A prototype of the IPT adapted for sample data will be demonstrated and participants encouraged to test it with their own data sets.

### 3. Information architecture – GEOSS perspective - Lorenzo Bigagli (GEOSS)

**Presentation:** Bigagli_slides.pdf

The session aims at introducing the architecture of the Global Earth Observation System of Systems (GEOSS), the GEOSS Common Infrastructure (GCI) and the GEOSS Brokering Framework. GEOSS has been created by an international voluntary effort that connects geospatial, Earth Observation and information infrastructures, acting as a gateway between producers of environmental data and end users. GEOSS aims at enhancing the relevance of Earth Observation and at offering public access to comprehensive, comprehensive, and sustained near-real time data, information and analyses of the environment. The GCI allows the user of Earth observations to access, search and use the data, information, tools and services available through GEOSS. The GEOSS Brokering Framework implements multi-disciplinary interoperability and lower entry barriers for both users and data providers, allowing them to continue using their tools and publishing their resources according to their standards.

The session includes a live, interactive demonstration of the GEOSS Discovery & Access Broker, based on material from the "Bringing GEOSS services into practice" workshop, which session attendants may practice on their own computer.

*Pre-requisites:*
Operating system: Windows, OS X, Linux
Memory (RAM): minimum 4 GB
Disk space: minimum 20 GB
Follow the instructions at: http://www.unige.ch/sig/enseignements/GeossInPractice/Start.html

**4. Data sharing and repositories in the GBIF network - Tim Robertson (GBIF)**

**Presentation**: eu-bon-gbif-repositories.pdf

The GBIF network is diverse, spanning more than 500 institutions and connecting thousands of databases using a variety of protocols and tools. The key components of the network include the data publishing repositories, a central coordinating registry and a sophisticated search index, which supports the GBIF portal - itself consider a data repository in wider networks. During this session a live demonstration of data sharing between repositories will be given, during which the architecture of the network will be described. An installation of the GBIF Integrated Publishing Toolkit (IPT) which acts as a data publishing repository will be used to demonstrate the services of the GBIF registry component, and specifically the management of data profiles (standards) available to data publishers. A dataset will be mapped, and registered with GBIF. Crawling components will be alerted automatically, and the data will be indexed and made available for discovery and access through the GBIF portal and web services API. Some observations about this architecture will be offered, including the opportunity to collaborate with the EU BON partners to improve data security through redundant storage.

This session is targeted for people interested in the GBIF architecture and key components, the data flows within the GBIF network, the GBIF publishing tool and those interested in interfacing with GBIF through the portal web services API. Being a live demo, opportunity will be given to address questions along the way, with the overarching goal that participants leave with a better understanding of the data flows than before the session.

**5. Data flow and modelling in virtual laboratory - Hannu Saarenmaa (BioVeL)**

**Presentation:** Hannu - lecture_GBIFfrance_1.pdf

**More information on** Ecological Niche Modeling workflows (ENM) and tutorial there in.

The Biodiversity Virtual e-Laboratory, BioVeL, addresses research challenges by having scientists and computer engineers working together to develop tools for pipelining data and analysis into efficient analytical pipelines, called "workflows." Workflows are complex digital data manipulations and modelling tasks that execute sequences of web services. BioVeL designs and deploys such workflows for a selected number of important areas in

systematic, ecological, and conservation research, e.g. for the analysis of data sets with ecological, taxonomic, phylogenetic, and environmental information.

BioVeL data refinement and ecological niche modelling workflows allow researchers to (i) explore, access, refine, and format large data sets from major data providers; (ii) combine disparate data sets with the researchers' own data; and (iii) run complex and computationally intense analytical cycles. (iv) generate comparative maps of species distribution.

The training workshop will demonstrate use of the informatics tools and services developed by the BioVeL project to address research topics such as historical analyses, invasive species distribution modelling, endangered species distribution modelling, and dynamic modelling of ecologically related species, and Essential Biodiversity Variables. In particular, there will be introduction to the BioVeL e-infrastructure and portal. Examples of taxonomic data cleaning, ecological niche modelling, model testing, statistical analysis of GIS data, invasive and endangered species distribution modelling, and historical comparison biodiversity from museum collections will be shown.

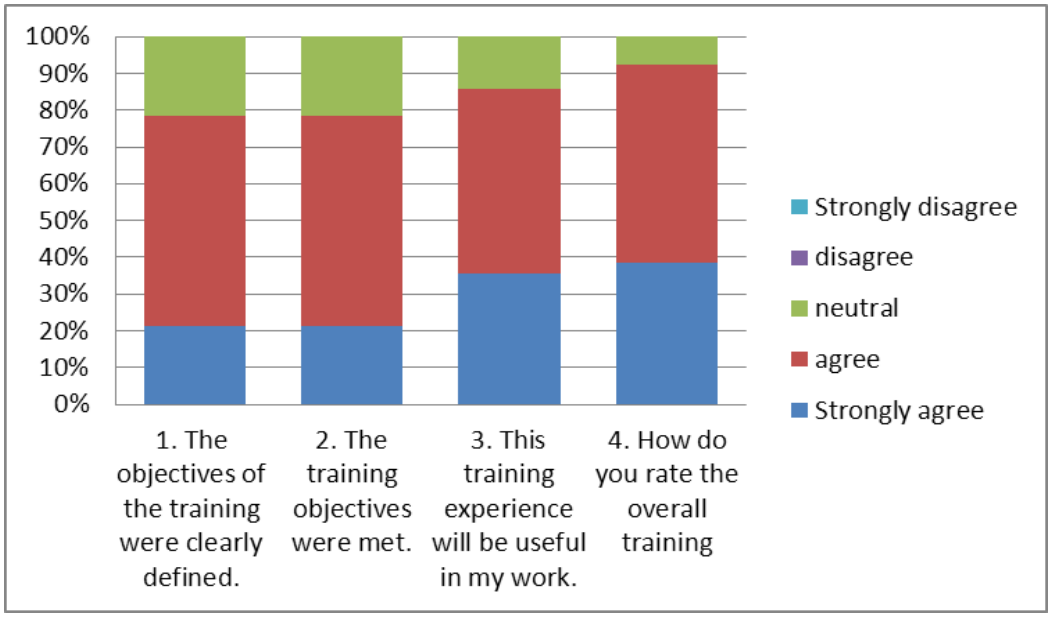
## 6. Data sharing and repositories in DataONE – Bruce E. Wilson (DataONE)

**Presentation**: DataONE Overview EU BON 2014-04.pdf

The mission of the Data Observation Network for Earth (DataONE) is to enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it. Organizations that collect, manage, or distribute data relevant to the Earth and the environmental sciences can collaborate with each other and with DataONE by becoming Member Nodes in DataONE. This collaboration brings broader exposure to the organization's holdings, tools for end-users to more directly access and use data (the DataONE Investigator Toolkit), and tools to assist the organization with their preservation and curation missions. DataONE also makes available a wide range of educational materials and best practice guides for community use in data management education and has conducted sociocultural studies on the barriers and enablers for improved data sharing. This talk will provide an overview of DataONE, highlight the sociocultural and technical approaches used by DataONE to enable data sharing and data interoperability, and explore ways that DataONE and other projects can collaborate with each other.

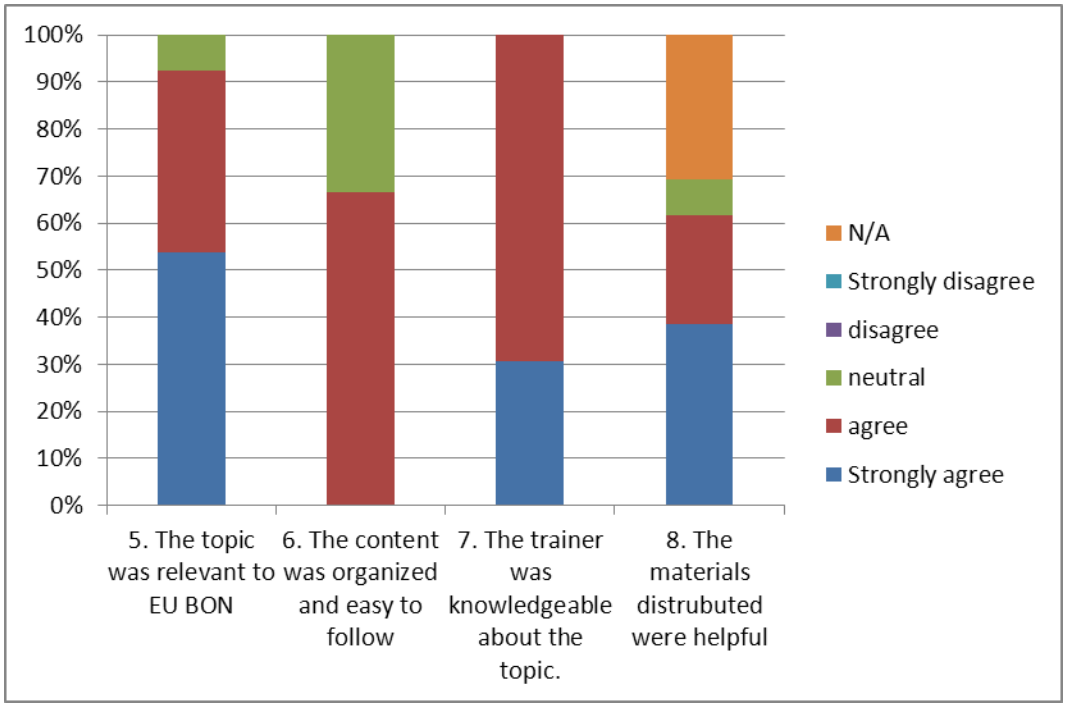
# **Annex 2:** Online feedback survey - results

### 1. Overall evaluation of the training:

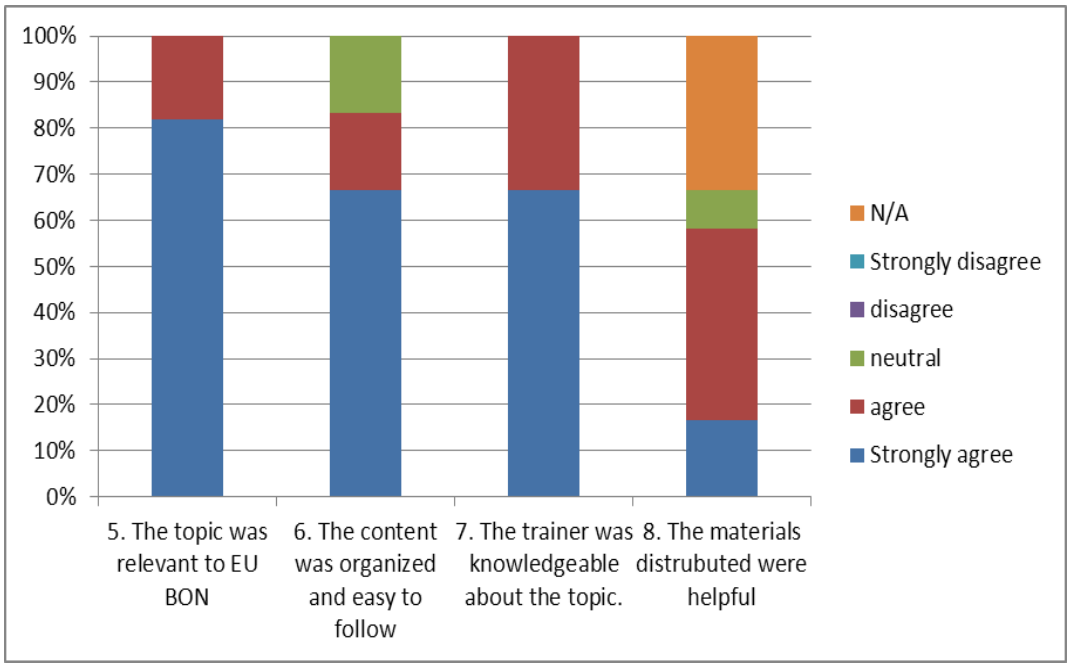



### 2. Evaluation per module:

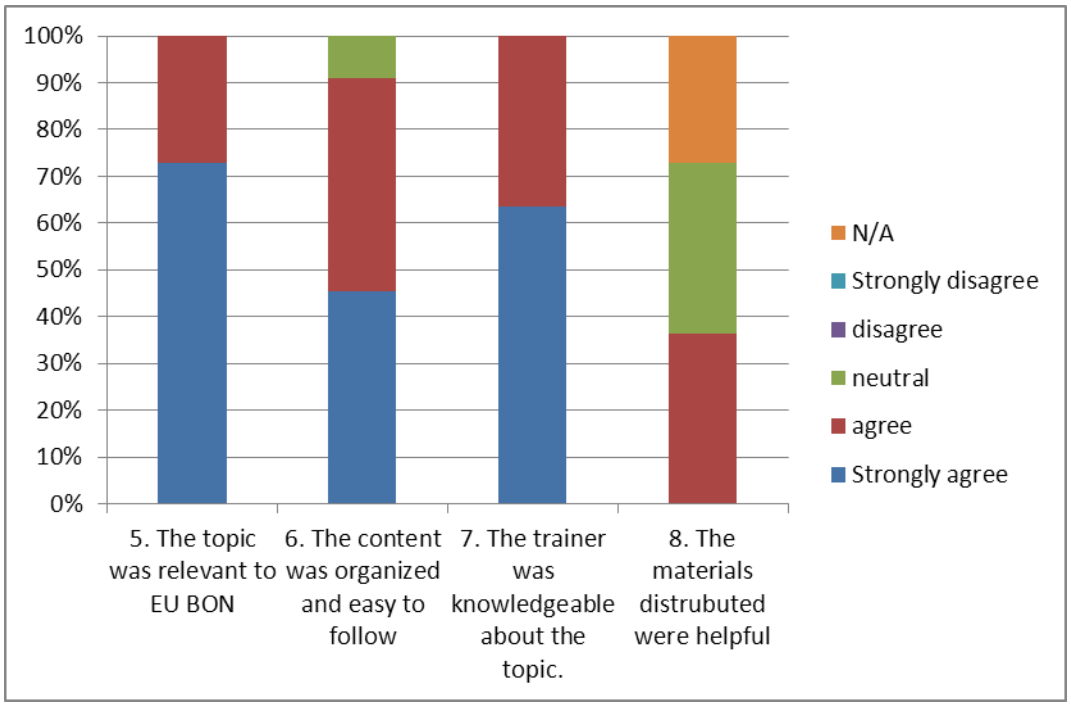
*Integrating biodiversity networks through Software Oriented Architecture (Antonio Garcia)*



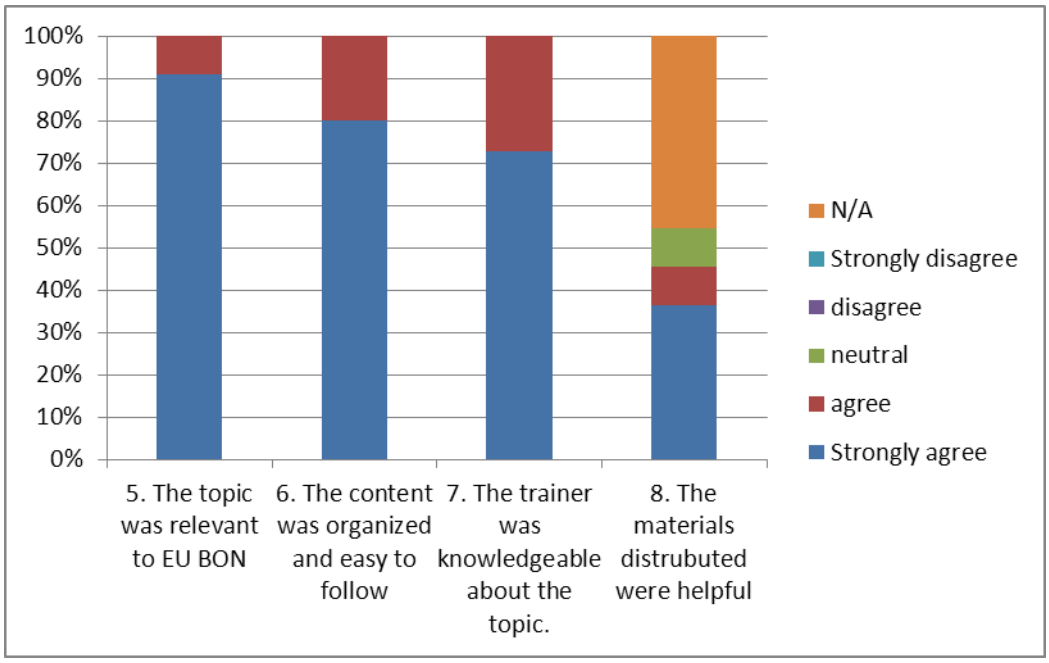
*Data standards: publishing sample-based data (Eamonn O Tuama)*



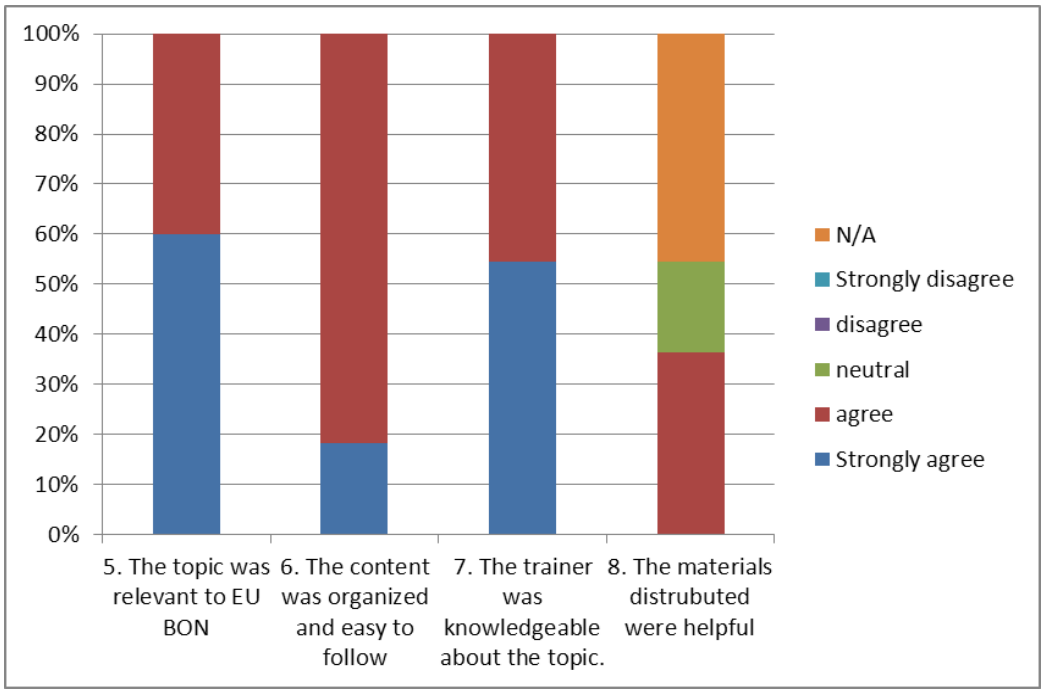

*Information architecture - GEOSS perspective (Lorenzo Bigagli)*



*Data sharing and repositories in GBIF network (Tim Robertson)*




*Data flow and modelling in virtual laboratory (Hannu Saarenmaa)*



*Data sharing and repositories in DataONE (Bruce E. Wilson)*