Annex 1: Definitions and Concepts

In order to evaluate and select appropriate tools for sharing data or metadata or any other data handling, a first need is a good understanding of what these terms mean and how they are used in life sciences. Within the context of biodiversity informatics one operates with terms like “data”, “standard”, “sharing”, etc., but their definitions may vary with the context and domain. To eliminate misunderstanding and potential misuse of the terms, fundamental concepts and definitions are first introduced.

Data

The many definitions and terms which include "Data" as part of their name, coined and documented in depth through numerous biodiversity infrastructures/interoperability projects, reflects the growing complexity in handling data flows and the increased need to formalize and categorize the multiple aspects of the notion of “data”. Furthermore, the integration of biodiversity data, which may include at least formats of genetic sequences, species occurrence (distribution/abundance/biomass/production) values and habitat maps, requires clear and unambiguous identifications of the terms for data.

Data is a set of values of quantitative measurement, or a qualitative fact on some entity in a structure of known format (e.g. spatial and tabular), typically the results of measurements. Data can be collected either automatically by devices like loggers or by individuals, which also determine the standards and formats in which they are produced. From these formats, information patterns and interrelations can be derived and subsequently interpreted, a process which provides evidence, which can, in turn, be used to create or enhance knowledge.

Data is often assembled in discrete units of digital content, such as files or records in a database, often expected to represent information obtained from a particular observation, sample, location, or period of time during a scientific study. These discrete units of data may be further organized into a dataset, which is an organizational tool to present a coherent and complete collection of data relevant to a particular topic. A dataset may be a single file or database, or it may be composed of thousands of files, and it is possible for a single database to contain many datasets. The organization of data into files and datasets is generally not standardized and depends on the particular needs of the individuals collecting the data and the anticipated uses of that data.

In the context of biodiversity observation network the term data should be associated with the purpose and the context in which this data is used whenever an ambiguous interpretation might arise.

Data standards

"Standards are documented agreements containing technical specifications or other precise criteria to be used consistently as rules, guidelines, or definitions of characteristics to ensure that materials, products, processes, and services are fit for their purpose"[1] (ISO 2015).

Data Standards are documented agreements aiming to provide consistent meaning to data shared among different information systems, programs, entities of data-consumers/users on representation, format, definition, structuring, tagging, transmission, manipulation, exchange, use, and management of data. Data standards in biodiversity science are being managed mainly by the Biodiversity Informatics Standards organization[2] TDWG.

Metadata

Metadata is "data about other data", based on standard specific to a particular discipline. Metadata is a description of content and context of content, using predefined attributes, aiming at providing a brief data about the characteristics of a resource (e.g. 'who, what, where, when, how and on what purpose').

In the GEOSS and GBIF contexts, from the point of view of the data provider, metadata contains information about its resources (datasets), while for the data consumer the metadata is used both to evaluate the resources and services needed to handle the data (e.g. discover, access) and to "assess appropriateness of the resource for particular needs – its so-called 'fitness for purpose'."[3]

Within the biodiversity domain the metadata description (file or data) should automatically be assigned to all processed and published data or object. Another requirement is that a tool for data sharing should guarantee a persistent link between the metadata and data/object. This is very important for the integrity of the information, to keep track of the origin of the data and respect IPR statements for example.

Depending on the context or usage, the same piece of information can be considered as metadata or data. The tools for data sharing can have embedded metadata templates, while in other cases the data standard is in part or entirely considered as metadata. Known standards that may fall under that case are for example Ecological Metadata Language (EML[4]), Darwin Core (DwC[5]), ISO 19115 (Geographic information – Metadata[6]) and Access to Biological Collection Data (ABCD[7]), to name a few. These and other data standards have been extensively reported in the EU BON deliverable D2.1 Architectural design, review and guidelines for using standards[8].

[1] http://www.iso.org/iso/home/standards.htm

[2] http://tdwg.org/

[3] https://code.google.com/archive/p/gbif-metadata/wikis/Introduction.wiki

[4] http://en.wikipedia.org/wiki/Ecological_Metadata_Language

[5] http://en.wikipedia.org/wiki/Darwin_Core_Archive

[6] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

[7] https://github.com/tdwg/abcd

[8] http://www.eubon.eu/documents/1/

Data vs. information

Data or “raw data” (also known as “primary data”) is a term for information collected from a source. From the perspective of the infrastructure service provider an important distinction between raw data and information is that data entities are provided, defined and described by an external source, which is outside of the scope of the infrastructure. Raw data is multi-purpose and can be reused. Raw data doesn't yield much information until it is processed (hence interpreted) and possibly integrated with other data. Once processed, the data may support particular types of information.

For example, an occurrence record for a certain species within a dataset is considered as "data". The interpreted contribution of one or a set of such records with its known attributes and relationships to other data, in term of scientific meaning, is "information".

The LifeWatch[9] information models, which aim to conform with the INSPIRE[10] Implementation Rules, address the differences between data and information (in accordance with Federal Standard 1037[11]) in its 'Information View'.

- Data: representation of measurements, facts, concepts, or instructions in a formalized manner that can be processed by humans or by automatic means.
- Information: the meaning that a human assigns to data by means of the known conventions used in their representation.

The LifeWatch Reference Model[12] further distinguishes between two aspects of information:

- Primary and derived information (including metadata) related to biodiversity data.
- Meta-information, that is: descriptive information about available information and resources with regard to a particular purpose (i.e. a particular mode of usage). Examples of 'Purposes of data' that are handled by different meta-information models include: Discovery, Orchestration, Collaboration, Identification, Authentication and Authorization, Provenance, Quality evaluation, Indexing, Retrieving, and Integration.

Processed and secondary data and information

Based on the increased availability of biological records, secondary information can be generated by processing and analyzing primary data using cutting-edge techniques for modelling, mapping, statistics, graphing and for visualization of data.

The non-exhaustive example products of secondary information and data products may include Red Lists, endangered species lists, observations that associate spatial coordinates, environmental data with habitat and landscape data, genetic data based on sequences and genes.

[9] http://www.lifewatch.eu/web/guest/home

[10] http://inspire.ec.europa.eu/

[11] http://en.wikipedia.org/wiki/Wikipedia:Federal_Standard_1037C_terms

[12] http://www.eubon.eu/getatt.php?filename=LW-RMV0.5_4310.pdf

The need for definition of data for purpose

The discovery, analysis, and interpretation of data, particularly for the purposes of generating information, often requires an understanding of the semantic context for a particular term, which depends on the particular scientific community and the purpose for which the data was collected. For example, precipitation has a very different meaning in the context of a chemistry dataset than an ecological dataset. And within ecology, the concepts of rain, snow, and sleet are understood to be specific forms of precipitation.

Ontologies are structured way to organize the different meanings that a particular term can have in different contexts as well as to describe the relationships between different concepts. Well-structured ontologies can greatly assist both the discovery and interoperability of datasets, but the proper application of these ontologies requires an understanding of the context of the data, which should be provided by the metadata. One mechanism of providing that information is to explicitly specify that context, by referencing a particular term in a relevant ontology or from a specifically referenced controlled vocabulary of keywords.

Some recent developments regarding vocabularies and ontologies in biodiversity informatics are outlined in deliverable D2.1.

Data publishing

Biodiversity data can be made publicly available through the process of “publishing”. Data publishing makes the data accessible through the use of standard procedures and protocols. It implies the use of common practices and standards ensuring that data can be discovered and reused effectively, and that data owners and custodians get the recognition they deserve. These practices also apply for data sharing, when data are made fully publicly available.

GBIF[13] and Pensoft[14] summarize the incentives to publish biodiversity data as follows:

- Data can be indexed and made discoverable, browsable and searchable through biodiversity infrastructures (e.g., GBIF, Dryad[15] and others):
- Discoverable and accessible data contributes to global knowledge about biodiversity, and thus to the solutions that will promote its conservation and sustainable use.
- Data publishing enables datasets held all over the world to be integrated, revealing new opportunities for collaboration among data owners and researchers.
- Publishing data enables individuals and institutions to be properly credited for their work to create and curate biodiversity data, by giving visibility to publishing institutions through good metadata authoring.

[13] http://www.gbif.org/publishingdata/summary
[14] http://www.pensoft.net/
[15] http://www.datadryad.org/

- Collection managers can trace usage and citations of digitized data published from their institutions and accessed through GBIF and similar infrastructures.
- Data produced and collected using public funds can be published, cited, used and re-used, either as separate datasets or collated with other data. Indeed, many funding agencies now require researchers to make their data freely accessible.

To encourage the publishing of biodiversity data one should stress the importance of the use of the 'Data papers' concept (recently promoted for the biodiversity community by Chavan and Penev (2011), Chavan *et al*. (2013).

A **data paper** is a searchable metadata document, describing a particular dataset or a group of datasets, published in the form of a peer-reviewed article in a scholarly journal. In contrast to the datasets published in conjunction with academic research papers, data papers may contain raw primary data, independent of a research hypothesis. This makes it uniquely adapted for the publication of biodiversity data from large collections, such as those curated by natural history museums.

Unlike a conventional research article, the primary purpose of a data paper is to describe data and the circumstances of their collection, rather than to report on hypotheses testing and to draw conclusions.

Key characteristics of the data-paper concept (Chavan *et al*., 2013) are that it:

- provides a citable journal publication that brings scholarly credit to data publishers;
- describes the data through structured, human-readable extended metadata;
- brings the existence of the data to the attention of the scholarly community.

Recent developments include the endorsement of the data paper concept by several EU-funded projects and the creation of the next-generation Biodiversity Data Journal[16]. Furthermore, Colombia's Alexander von Humboldt Biological Resources and Research Institute is commissioning a journal dedicated to publishing data papers, and public repositories, such as Dryad and Scratchpads, are collaborating with academic publishers to encourage data-paper publishing (Chavan *et al*., 2013).

Data sharing and open access

Wikipedia defines data sharing as "the practice of making data used for scholarly research available to other investigators"[17]. It's considered to be a part of scientific method together with documentation and archiving. A number of institutions, funding and publishing agencies have policies regarding data sharing. While data sharing for some is about validating results, for others, publishing data are about enabling big data solutions and approaches (Anderson, 2014).

[16] http://bdj.pensoft.net/
[17] http://en.wikipedia.org/wiki/Data_sharing

The terms “data sharing” and “data publishing” are often used interchangeably. However, there are differences. Data that is shared may still be private and access to it can be controlled. Access to shared data can be revoked. (This was an important clause in the original GBIF Data Sharing Agreement, which placed emphasis in keeping the data owner in control). However, when something is published, it has been made openly available for good, and access cannot be revoked anymore.

Shared data are useful only if they are searchable and usable. For both characteristics data must be formatted in a standard way, conform to standard structure and semantics and have appropriate metadata attached[18].

Despite the ongoing discussion how to share, what to share and on what conditions to share it’s almost impossible to imagine the modern science without data sharing initiatives emerging worldwide and in different disciplines.

Open access is an important principle in data sharing (although data can also be shared in restricted ways). Data sharing necessitates the use of an agreement or a license where the terms and conditions have been stated. When integrating data from thousands of sources, only open access and standardized licenses such as those of Creative Commons may work.

The important players in domains of earth and biodiversity observation, such as GEO BON, GEOSS, including EU BON, pursue strategic goals[19], among which data sharing is directly addressed:

- address the need for timely, global and open data sharing across borders and disciplines, within the framework of national policies and international obligations, to maximize the value and benefit of Earth observation investments,
- implement interoperability amongst observational, modelling, data assimilation and prediction systems.

The first 10-Year Implementation Plan of GEO stated that "The societal benefits of Earth observations cannot be achieved without data sharing", and set out the GEOSS Data Sharing Principles:[20]

- There will be **full and open exchange** of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products being provided free of charge or no more than cost of reproduction will be encouraged for research and education.

[18] http://www.nature.com/nature/journal/v461/n7261/full/461171a.html
[19] https://www.earthobservations.org/documents/geo_vi/12_GEOSS%20Strategic%20Targets%20Rev1.pdf
[20] https://www.earthobservations.org/geoss_dsp.shtml

EU BON Data Sharing Agreement

The EU BON project determined in 2013 the need to put in place a detailed Data Sharing Agreement[21], which follows the above GEOSS Data Sharing Principles, but also gives additional terms and conditions, which are relevant for the biodiversity community. These conditions include the need to hide potentially sensitive data on endangered species, and the need for an embargo on data release to support priority in scientific publishing, and to motivate data sharing. This agreement has yet to be tested in practical terms.

Other related initiatives include the revision of the GBIF Data Sharing Agreement to ensure that all data sets are associated with a standard, machine-readable Creative Commons equivalent license (i.e. CC-0, CC-BY, CC-BY-NC) that can be automatically processed to support data integration across large number of datasets, and the Bouchout declaration[22] that promotes licenses or waivers in support of open biodiversity knowledge management. The EU BON Data Sharing Agreement is in line with the main principles of the Bouchout declaration on open biodiversity knowledge management. Recommendations that are beyond the scope of the agreement are also promoted (e.g. the need for persistent identifiers for data, linking data using agreed vocabularies and sustaining identifiers in the long term) (Wetzel *et al*., 2015).

Moreover, EU BON adheres to the principles of free and open exchange of data and knowledge, in accordance with the “Joint Declaration on Open Science for the 21st Century”, presented by the European Federation of Academies of Sciences and Humanities and the European Commission on 11th April, 2012[23].

References

Anderson K (2104) Data sharing and science — Contemplating the value of empiricism, the problem of bias, and the threats to privacy. http://scholarlykitchen.sspnet.org/2014/03/05/data-sharing-and-science-contemplating-the-value-of-empiricism-the-problem-of-bias-and-the-threats-to-privacy/ (visited 11/04/2016)

Chavan V., Penev L. (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science . BMC Bioinformatics, 12: S2; . DOI: 10.1186/1471-2105-12-s15-s2

Chavan V., Penev L., Hobern, D. (2013) Cultural Change in Data Publishing Is Essential. BioScience 63(6): 419‑420. DOI: 10.1525/bio.2013.63.6.3

Wetzel FT, Saarenmaa H, Regan E, Martin CS, Mergen P, Smirnova L, Ó Tuama É, García Camacho FA, Hoffmann A, Vohland K, Häuser CL (2015) The roles and contributions of Biodiversity Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case study, Biodiversity, DOI: 10.1080/14888386.2015.1075902

[21] http://www.eubon.eu/news/10954_EU%20BON%20Data%20Sharing%20Agreement

[22] http://www.bouchoutdeclaration.org/declaration/

[23] http://www.allea.org/Content/ALLEA/General%20Assemblies/General%20Assembly%202012/Joint%20Declaration%20GA%20Rome%202012%20signed%20v2.pdf