Machine-Readable Data Formats in Biodiversity Informatics

Recommendations and Best Practices

Project idea: Giuditta Parolini, Data Scientist, Museum für Naturkunde Berlin, Germany

Table of Contents

- 1. Preamble
- 2. Introduction
- 3. What does machine-readable mean?
- 4. Why do machine-readable data matter for biodiversity science?
- 5. General formats for machine-readable data
  - 5.1. Tabular data
    - CSV, TSV
    - TXT
    - XML
    - JSON
    - RDF
    - Parquet
  - 5.2. Geographic data
    - Shapefile
    - GeoJSON
    - GML
    - GeoTIFF
    - GeoPackage
  - 5.3. Sequencing data
  - 5.4. Images
  - 5.5. Other media
- 6. Formats for machine-readable data specific to biodiversity science
  - DwC-A

    - ABCD and extensions
    - Other machine-readable formats used in biodiversity science
- 7. Machine-readable data in the age of AI: Some final remarks
- 8. References
- 9. Acknowledgements

1. Preamble

This Guide on machine-readable data formats is written in Markdown as an invitation to start exploring tools with higher potential for machine-readability and information sharing compared to standard word processors. Markdown is a markup language. A document in Markdown format can be opened and edited using any text editor and does not require dedicated software like a PDF or a DOCX file. The trade-off is that styling options are limited compared to standard word processors, but human-readability remains satisfactory.

If you wish to see the rendered Markdown document, there are many online and offline open-source editors available (e.g., VSCode + Markdown extensions). These editors are all largely compatible, but Markdown has many flavors and different editors may render some elements of this document in a slightly different fashion. A few Markdown editors do not support HTML style elements. This will prevent you from accessing the internal links in the sections *Table of Contents* and *References*. If possible, choose a Markdown editor that also supports HTML.

If you prefer to consult this guide in a more human-readable format, this is not an issue. A Markdown document can be turned into a PDF in a few clicks. This document can even be published to a webpage, as the transformation in HTML is equally straightforward. You can then decide to read this document as a human reader would, have a computer extract automatically pieces of information that are of special interest to you, even scrape the data when this document is publicly accessible on a webpage.

Markdown integrates seamlessly with version control systems and can, therefore, be used for collaborative projects. This Guide is published as a GitHub repository to give multiple authors the opportunity to contribute. If you wish to improve this document or you have comments and suggestions after reading it, you are welcome to create an issue in GitHub or make a Pull request.

The Guide on machine-readable data formats is complemented by a **Jupyter notebook with examples**. The notebook and related data are in a separate folder in this repository. I suggest to go through the examples while reading this document, as the cases selected give a clear illustration of the statements made. In order to access the notebook, you need a local installation of Jupyter notebook or a cloud service that allows access to Jupyter notebook. All the data for running the examples are available in the folder.

Alongside this materials, the GitHub repository also includes a **glossary** defining key terms used in the Guide, such as *data format* and *data standard*, and a quick reference **list of open data formats** divided according to data type (tabular, geographic, etc.).

2. Introduction

This overview of recommendations and best practices for creating and sharing machine-readable data in biodiversity informatics wishes to contribute to the development of FAIR and sustainable data practices. The availability of machine-readable data is a key usability requirement, as it enables data sharing, (re)use, and integration at scale, and provides opportunities for automated quality control.

Paradoxical as it may sound, however, some of the most popular file formats, e.g., PDF, are not suitable for creating and transferring machine-readable data. For instance, tabular data embedded in a PDF file such as a journal article, are human-readable, but cannot be accessed directly by a computer, while the same data presented in a properly formatted CSV file are machine-readable.
Another interesting example is the DOCX format, which is very popular among word processors. The DOCX format is based on the standard Office Open XML which was defined to foster "interoperability across office productivity applications and line-of-business systems, as well as to support and strengthen document archival and preservation, all in a way that is fully compatible with the existing corpus of Microsoft Office documents" (ISO/IEC, 2016). In principle, a DOCX file can be easily parsed being based on the XML standard, in practice the same example described above, i.e., the extraction of a data table from an article manuscript in DOCX format, requires labour, starting with accessing the multiple files which are hidden in the DOCX file, recovering then the table-only content in the file, and eventually exporting the table data in a machine-readable format like CSV for further use.

As biodiversity science fully endorses data sharing and openness, an additional consideration worth making is that large part of automated information sharing takes place via the Internet nowadays. Data can be transferred and distributed via web APIs using XML and JSON format, but data can also be scraped directly from webpages, if permitted by intellectual property laws and regulations. Also in

this case, however, the data table in the example mentioned above is not likely to be easily machine-readable. Files in PDF and DOCX format can be transformed in HTML for web publication, but the results are usually poor when complex styling options are used in the original file, as it is often the case with word processors. A web scraping effort would therefore be laborious and error-prone.

In essence, therefore, before choosing a data format due consideration should always be given to machine readability. Just very small datasets with at most a few hundred entries can be processed directly by humans. For all the other datasets, machine processing is the only viable solution, but computers can work with data only when the requirement of machine-readability is satisfied. Providing a dataset in a non machine-readable format forces the user to laborious preprocessing and makes the chance to introduce mistakes in the original data a real issue.

Non machine-readable datasets run the risk to be forgotten and ignored regardless of their scientific value. The aim of this guide is to clarify the importance of producing machine-readable data and help data managers and biodiversity researchers select the best machine-readable format according to their needs.

The document begins presenting the concept of machine-readable data and why it is especially valuable in biodiversity informatics. The core part of the Guide is filled with suggestions and recommendations for machine-readable data formats suitable for biodiversity informatics. Both machine-readable data formats of general use, like CSV, and machine-readable data formats that are specific to biodiversity science, like DwC-A (Darwin Core Archive), are discussed. A final section will sketch a few considerations about the value (and pitfalls) of machine-readable data in the age of AI.

3. What does machine-readable mean?

The concept of machine-readable data has a long history. It was introduced in the 1960s to refer to the physical input formats, such as magnetic tape and punched cards, typical for computers at the time, and was at the bottom of the development of the MARC (Machine-Readable Cataloging) programme initiated by the Library of Congress in the US to automate library operations (Avram, 2003). Nowadays, the concept of machine-readability has shifted from the realm of the physical media used to feed data into a computer to the choice of a suitable digital format for automatic data processing.

The Open, Public, Electronic, and Necessary Government Data Act (2017-2018), which was passed by the US Congress, defines a machine-readable data format as "a format in which information or data can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost" (OPEN Government Data Act §3561(7)). The European Union takes its definition of

machine-readability from the Open Data Handbook. Here, a machine-readable data format is defined as a structured data format that can be automatically read and processed by a computer. The definitions cited can be considered equivalent as the existence of a structured data format is a pre-requisite for maintaining the semantic meaning while automatically processing the data. A CSV document is an example of structured data because, for each data point (each row of the CSV file), a certain amount of information, i.e., the column values, is provided.

It is important to point out that the file extension, e.g. .csv, is a necessary, but not a sufficient condition to have machine-readable data. It is necessary, because the data format must be structured. Not only CSV, but also XML, JSON, etc., fulfil the requirement of being structured data formats and the choice of one over the other depends on the type of data examined and the specific use case. The file extension, however, is not sufficient, because a badly formatted spreadsheet with headers on multiple rows, deviations from the tabular format, improper use of the column delimiters, etc. cannot be turned into a machine-readable file just by saving the spreadsheet in CSV format. The CSV will be invalid and any attempt to read it, regardless of the software tool used, will result in an error message.

It should be also evident that **electronic access to a document does not mean that the document is machine-readable**. For instance, coming back to the example cited in the introduction, a PDF version of an originally printed journal article is not likely to be machine-readable. This becomes quickly evident by looking at the website of a scientific publisher with a notable tradition such as the Royal Society. PDF copies of scanned images of the print-only articles are displayed online, but the only piece of text available on the webpage is the abstract. Web crawlers are authorised to access it, but readers are warned that the abstract can contain typos, as it has been automatically harvested from scanned images of the printed version using Optical Character Recognition (OCR). Aside from the abstract, the Royal Society has made no attempt to make the data contained in its older articles. Indeed, once data have been produced in a non machine-readable format, their transformation into a machine-readable format is always an expensive (time, money, and human resources) and complex task that leaves plenty of room for errors and can considerably lower data quality.

In summary, **machine-readable data only** need to satisfy the criteria mentioned above, i.e., they **need to be digital (necessary, but not sufficient condition) and to use a valid structured data format that conserves the semantic meaning**. These requirements have some interesting consequences that merit more attention than they usually receive.

a) **Machine-readability is not defined in contrast or opposition to human-readability.** Unsurprisingly, therefore, machine-readable data are in many cases also human-readable (notable exceptions are binary files). Data in CSV, XML, or JSON format, for instance, can be inspected visually and many code editors can greatly enhance human-readability for these machine-readable

formats using indentation and different colours according to the components of the data structure. The fact that machine-readable data formats can be rendered on the screen in an appealing style, though, should not be confused and equated with the opportunity to style tables and documents offered by word processors and electronic spreadsheet software generating DOCX and XLSX files. While in the former example, the styling is superimposed at the display stage to make the information more appealing to the human eye, in the second case, instead, the style elements (for instance text boxes in a spreadsheet) are intrinsic to the construction of the file and they constitute an hindrance to machine-readability.

b) **Images are typically not machine-readable.** With the notable exception of barcodes and watermarks **[1]**, images are not machine-readable because the information they provide does not follow a specific data structure. This is true for holiday photos as well as for scientific images produced by microscopes, CT scans, infrared cameras, and other equipment commonly used in research. Images produced with digital equipment, though, usually can (and should) have associated metadata in machine-readable format. In addition, although the semantic meaning of images cannot be directly accessed by machines, computer vision algorithms for image segmentation and object detection have been in use for quite some time to facilitate information extraction from images and their automatic labelling.

[1] Barcodes and watermarks are image-based data interfaces. The information they encode can be recovered from their representation, for instance in print, via a sensor such as the camera of a mobile phone. Popular barcodes are the universal product code (UPC), the quick response (QR) code, the Aztek code, and the Data Matrix (Sharma, 2016). While barcodes work by embedding the data in the image pattern, watermarks hide the embedded data in the image.

c) **There are potentially no limitations to the complexity of the data structures in machine-readable formats like CSV, XML, or JSON, nor these data formats have limitations to the number of data points they can support.** On the contrary, Excel XLSX files (but the same is true for other proprietary spreadsheet software) have limitations in both the number of columns and rows they support. The limit was set by the developers of this proprietary software tools to avoid overloading memory and computing capabilities of the average user. Also for machine-readable data formats like CSV, XML, or JSON, however, parsing and editing larger and larger data files takes longer and longer until computational and memory capacity of the hardware used are reached.

4. Why do machine-readable data matter for biodiversity science?

At the time of writing (Summer 2024) the Global Biodiversity Information Facility GBIF makes available via its partner institutions over 100.000 datasets amounting to several millions data records on plants, animals, and natural environments. These data range from taxonomic checklists to species observations to images and other media of digitised natural history collections.

While the raw data in themselves might not be machine-readable (this is the case for the digital images of collection specimens), all metadata for the records and all the tabular data are made available in a machine-readable format following one of the data standards popular in biodiversity science. This makes all the records searchable and retrievable not just manually, but also automatically using the GBIF API.

Given the amount of records under examination, it becomes quickly apparent that having available information in a machine-readable format offers the opportunity to access data at scale, combine data for analysis that goes beyond the original aims of the data creators, and redistribute data to a larger public. Yet, this is not always the case, especially with the smaller datasets associated to scientific articles. Due to the lack of a general policy for biodiversity science publications and the fact that biodiversity research appears in journals belonging to different disciplines, it is not uncommon to still find published biodiversity data only available in tables saved in DOCX files or compiled in spreadsheets that are not machine-readable. An even greater concern is for the unpublished data produced during research projects. The increased availability and usage of data management tools and platforms facilitate the implementation of best practices, but these tools are not ***per se*** a guarantee that the data created are machine-readable data.

For this reason, this guide provides an overview of suitable machine-readable data formats for biodiversity research, illustrates their context of application, and gives advice against possible pitfalls in their implementation.

5. General formats for machine-readable data

General formats is here used to refer to all machine-readable data formats that do not have disciplinary boundaries, but are widely adopted in science, business, and administration. There are quite a few of them. This section will not survey them all, but it will provide an overview of the most popular formats according to possible use cases.

Image data and other media data are included in this section, even though these data are unstructured and therefore, strictly-speaking, they are not machine-readable data. The reason to

include them in this guide, however, is twofold. The first reason is that these data should at least have associated metadata (time, place, creator, rights, etc.) that are machine-readable and that provide relevant information about data creation and use (for instance, the license under which these data are available). The second reason is that unstructured data are becoming a big chunk of the data currently produced in biodiversity science. For instance, images, ranging from photos collected by citizen science projects to high-resolution specialised microscopy images, are becoming more and more relevant in biodiversity. Extracting the information they contain is an increasingly important pursuit that deserves some attention, as the methods applied are automatic methods that can generate machine-readable data for further analysis.

5.1. Tabular data

Several types of scientific data, such as measurements and observations, are tabular data, that is to say, they are organised in rows and columns containing numbers, text strings for categorical/non-categorical data, urls, etc. These tabular data can be created using a spreadsheet software or extracted from a database via a query language. Regardless of how they are generated, tabular data should always be made available in a machine-readable format that makes their analysis and re-use immediate. Many opportunities are available.

- **CSV, TSV**

  CSV (Comma Separated Values) files (extension .csv) display tabular data using the comma (“,”) or the semicolon (“;”) as a delimiter. The semicolon as a delimiter is popular in European countries, where the comma is generally used as a decimal separator. For standard tabular data containing figures and text strings and with a number of rows up to about one million **[2]**, the CSV data format should be preferred because it is an open format and its portability is much higher compared to spreadsheet formats like .xlsx or .ods. Data in CSV format can be easily inspected and analysed using popular programming languages for data science, like Python and R.

  TSV (Tab Separated Values) files differ from CSV files only for the choice of the delimiter that is the tab in this case. One of the two formats made available by GBIF for data download is precisely the TSV file format.

  CSV/TSV files have a very simple structure that can be easily inspected by eye. The first row contains the column headers and all the other rows in the file are the data points included in the dataset. The presence of the column headers is not a validity requirement for a CSV/TSV file. Only the data rows may be present. It is, however, required that the delimiter used in the file - be

it the comma, the semicolon, or the tab - is never used in the value fields, unless properly escaped or marked as a string. If this requirement is not satisfied, the file cannot be read properly and a tokenization error (i.e., splitting the data into columns) will be raised.

A limitation in using the CSV data format is that it is not possible to add directly any comment (containing, for instance, metadata such as data author or data license) to the file. *Ad hoc* solutions for adding comments to a CSV file will not be recognised by standard spreadsheet software and may cause issues when the data are read. A general recommendation is to add a README file (usually a Markdown or a TXT file, but XML or JSON files would also be suitable) to the CSV file and include in the README the information that would have been placed in the comments.

Recommendations

- If the CSV file use the semicolon as a separator, it is advisable to convey this piece of information in the README file and/or other sources describing the dataset.
- Data creators using the CSV data format should always checked whether their CSV files are valid. Forgotten and not properly escaped commas, semicolons, and tabs in data cells are recurrent causes of tokenization errors.
- Saving in CSV format a badly formed spreadsheet (e.g., column headings split across multiple lines, multiple tables on the same electronic sheet, etc.) will not make the data machine-readable. Tabular data should be arranged with one table per sheet, column headings (if any) in single cells on the first row of the spreadsheet, no white spaces in between the data rows, etc. to create a valid machine-readable CSV file from the original spreadsheet.
- If CSV files contain textual information, it is always recommended to properly encode this information (encoding in UTF-8 is usually recommended) to avoid issues with non-standard characters when redistributing the data files.

[2] One million rows is here suggested because many spreadsheet software tools, e.g. Excel, have limitations in opening CSV files larger than this. However, CSV files with several million rows exist and can be managed using different software tools (e.g., a text editor). The CSV data format, however, is not best suited for very big datasets for which other formats, like Parquet described below, are a better choice.

- **TXT**

TXT files are machine-readable, but only support plain text format. UTF-8 is the recommended character encoding method to avoid issues, similarly to what already suggested for CSV/TSV

files. TXT files should be the preferred machine-readable format for unstructured text that needs to be further analysed/mined. For annotated text, a machine-readable format like XML could be used. Tabular data can also be manipulated as strings of characters, even when the string characters are all numbers, therefore TXT files can be used to store tabular data. However, the CSV/TSV file format should still be the preferred choice in this case.

- **XML**

  The eXtensible Markup Language (XML) is a file format used for storing and transferring data. It is called eXtensible Markup Language, because the content in a XML file is a combination of tags, which logically structure the content, and proper data. XML data files are both machine-readable and human-readable. The XML data format was developed in the context of the World Wide Web Consortium (W3C) and it is a W3C recommendation, i.e., a standard for web communication. All XML processors accept the character encodings UTF-8 and UTF-16. UTF-8 is also in this case the recommended encoding to avoid issues at reading time.

  If the XML document has a **prolog** (optional feature) [e.g., <?xml version="1.0" encoding="UTF-8"?>], it must be placed at the beginning, before the file content. Every XML file has a **root** entity, which is the entity a parser tool uses to begin reading the content of the XML document. All the content in the XML file is packaged in storage units, the elements, enclosed by appropriate start- and end- tags and properly nested. Tags in a XML file are case sensitive and cannot contain empty spaces. The only reserved tag, not available to the xml document creator, is xml itself.

  XML documents must be well-formed according to the WC3 specification, otherwise a fatal error is raised by the parser. In addition, a distinctive feature of the XML file format is that it can be also validated using either a **XML Schema Definition (XSD)** or a **Document Type Definition (DTD)** and an appropriate parser and it can also have comments, unlike CSV files. XML elements can have attributes. For instance, the literary genre can be an attribute for a book element in a XML file that lists the contents of a library. Attributes are enclosed in element tags and need to be put within quotation marks.

- **JSON**

  The JavaScript Object Notation (JSON) format is a data interchange format that is both machine-readable and human-readable. As suggested by the name, the JSON format is built upon the JavaScript syntax and standard. However, JSON is language independent and JSON files can be created and processed using any programming language of choice because the structures used

by JSON are common to all the main object oriented languages.

A JSON data file consists of **objects** included within curly braces, **data** assigned as key-value pairs included within quotation marks if strings, left unquoted if numbers, and **arrays** included in square brackets. The various data elements in a JSON file are separated by **commas**. UTF-8 is the recommended character encoding also for JSON files.

JSON is probably the most popular data exchange format for web APIs, as it is a lightweight alternative to XML for compact data serialization. JSON documents are also increasingly popular for use in NoSQL databases like MongoDB because they are flexible in storing multiple data types and can easily accommodate changes in data model as information is saved as documents rather than as tables. The only condition is that the JSON documents satisfy all formal validity requirements for the JSON format (e.g., no curly or square braces left open, proper use of the comma to separate elements, etc.)

- **RDF**

  The Resource Description Framework (RDF) is a W3C standard for data interchange on the web. RDF files allow to exchange data using a triplet syntax (subject-predicate-object) that can be turned into a directed graph. Each triplet element is identified by a URI (Uniform Resource Identifier). Triplets can be serialised using multiple formats. A couple of the most popular serialisation formats are RDF/XML (.rdf) and JSON-LD (.jsonld).

- **PARQUET**

  The PARQUET file format (file extension .parquet) is machine-readable, free, and open-source. Like JSON it is a language agnostic format that can be manipulated using multiple programming languages. PARQUET can be used to store tabular data, but also images, videos, and documents.

  PARQUET files store data column-based rather than row-based, as in a CSV. This means a reduction in memory space occupied by the data and faster query execution due to data skipping and predicate push down (using metadata with summary statistics to ignore irrelevant data objects at query time). In addition, the PARQUET file format can be efficiently compressed and decompressed and it is suitable for storing complex and nested data structures. Due to these features, PARQUET files are popular for accessing and querying datasets in the GigaByte range.

5.2. Geographic data

Geospatial information is popular in biodiversity science. It is employed to assess areas with high biodiversity and track changes in species distribution over time, just to mention a few of the possible uses. This type of information is crucial for informing species conservation and management strategies, and biodiversity data portals, like GBIF **[3]**, routinely make available tools to overlay the datasets they are indexing on geographic maps.

Geographic data deserve, therefore, a section of their own in a guide on machine-readable data formats for biodiversity science. The good news is that **proper geospatial information data formats, such as GeoJSON or GeoTIFF, are machine-readable**. The not so good news is that there is still geographic and location information distributed using image formats without proper georeferencing and this makes (re)using the geospatial information in the images problematic.

There are dozens of geospatial information data formats available, both proprietary and non-proprietary, and it will not be possible to examine them individually in this guide. In the following section only a few of the formats most popular in biodiversity science will be considered. For a general overview of geospatial data formats based on open standards, please consult the website of the Open Geospatial Consortium.

Geospatial data can be split in two main categories: vector and raster data.
Geospatial vector data consist in discrete vertices (or points) that create spatial objects. The vertices are defined on a two-dimensional coordinate system and, depending on how they are connected, they can have one of the following geometries: points (individual vectors), lines (at least two connected vectors per line), polygons (at least three connected and closed vectors per polygon). Attributes can be assigned to each geometric feature in geospatial vector data (e.g., a biodiversity index can be attached to the polygons defining a geographic area). Geospatial vector data can be quite efficient for data storage and management.

Geospatial raster data consist, instead, of a matrix of pixels organized into a grid, similarly to the representation of digital images. However, unlike standard images, in geospatial raster data each pixel is associated to a geographic location and can have information (continuous or categorical) attached to it (e.g., land elevation).

In many cases, machine-readable geospatial data formats are built on machine-readable data formats discussed above, like XML and JSON, that are customised to accommodate for geographic information for specific use cases. Two notable examples are Open Street Map XML that is the export format for the information contained in the Open Street Map geographic database and Google KML, an XML-based file format introduced specifically to display geographic information on Google Earth.

[3] GBIF Maps API is a web map tile service based on an open standard of the Open Geospatial Consortium (OGC), an organisation devoted to improve access to geospatial and location information.

- **Shapefile**

  The Shapefile is a spatial vector data format introduced by ESRI, the company developing popular software for geospatial data like ArcGIS. Although the Shapefile was developed as a proprietary format, ESRI has released technical documentation that allows anyone to develop software tools, like the open-source software QGIS, for using Shapefiles.

  Shapefiles do not carry information in a single file, but in a set of different file types stored in the same directory and providing indications on geometric shapes and data attributes. The three mandatory components of a Shapefile are a **.shp** file that gives the shape geometry, a **.shx** file that gives the shape index position, and a **.dbf** file that contains the attribute data.

  Due to the long usage history and the relative ease to manipulate them, Shapefiles remain one of the most popular formats for transferring geospatial information. But they are not always the best choice. For instance, they have limitations in the amount of data they can accommodate (roughly 70 million data points), they use inefficient compression algorithms that make the data files bulky, the dBASE files (.dbf) that store data attributes do not support null values.

- **GeoJSON**

  GeoJSON is a spatial vector data format for encoding geospatial information and related non-spatial attributes based on the JSON data format presented above. A GeoJSON file must follow the open standard specification RFC7946. Coordinates in a GeoJSON file are given with reference to the coordinate system World Geodetic System 1984 (WSG84) only, as the use of alternative systems in the past proved to cause compatibility issues.

  GeoJSON objects can represent points, curves, and surfaces in coordinate space (Geometry), a spatially bounded entity (Feature), or a list of Features (FeatureCollection). Supported geometries in GeoJSON are ***Point***, ***LineString***, ***Polygon***, ***MultiPoint***, ***MultiLineString***, ***MultiPolygon***, and ***GeometryCollection***. Features are a combination of geometries and additional properties. According to the standard specification (see Introduction), GeoJSON "is concerned with geographic data in the broadest sense; anything with qualities that are bounded in geographical space might be a Feature whether or not it is a physical structure". For instance, GeoJSON is used as the data exchange format for annotations on digitised images of glass slides by the

open-source software QuPath.

Compared to the Shapefile, GeoJSON is a lighter data exchange format and, like JSON, it is well suited for transferring data to and from web APIs. Web development was indeed the main reason behind the development of the GeoJSON data format.

- **GML**

  The Geography Markup Language data format (file extension .gml or .xml) is based on the ISO Standard ISO 19136-1:2020 that has been jointly developed by the ISO and the Open Geospatial Consortium. This ISO standard defines an XML encoding for transferring and storing geographic information conveying both spatial and non-spatial (data attributes) properties related to geographic entities.

  As already described above for XML, GML allows for standardisation and verification of the transferred information. GML supports all common geometries used in geographic information systems like points, lines, and polygons. In addition, it can be used to store and transfer more complex information on coverage, topology, and time series. Being based on the XML standard - remember that the X stays for ***eXtensible***- GML schemas can be easily extended to accommodate a new piece of information.

- **GeoTIFF**

  The Geographic Tagged Image File Format (GeoTIFF) is a raster data format used to store and transfer image files (e.g., satellite images) that also contain contextual geographic information in the form of metadata tags. Typical information contained in the metadata tags are the coordinate system used, the projection, etc. In a nutshell, the metadata provide all the information required to automatically overlay the image to a coordinate map. The format is based on the TIFF file format and the file extension for GeoTIFF images is simply .tif.

  However, while the geographic information is in the form of machine-readable data, the image itself remain an unstructured data type whose semantic content cannot be extracted automatically, unless *ad hoc* machine-readable labels declaring position and nature of relevant objects (e.g., rivers) are added to the image in the form of annotations. For an overview of possible ways to extract information from GeoTIFF files please consult the section on images in this document. GeoTIFF images can be very large and bulky as lossless TIFF files preserve image quality, but at the expense of image size.

- **GeoPackage**

  GeoPackage is a data format for storing and transferring geospatial information using a SQLite container. GeoPackage is based on an open standard that defines how and what kind of information can be stored in the container.

  GeoPackage files have the extension .gpkg and can contain both vector and raster data. Geospatial data in GeoPackage format can be efficiently transmitted because SQLite databases are lightweight and serverless and all the information contained by the databases and their schema can be transferred in a single file.

5.3 Sequencing data

Sequencing data are increasingly employed in biodiversity science. They can be used to assess biodiversity, inform conservation efforts, and plan biodiversity restoration actions (Theissinger et al., 2023). All data formats currently in use for storing and transmitting sequencing data are machine-readable. These data, in fact, are generated as part of automatic workflows in which the sequencing information is expected to be consumed by machines and not by humans.

FastA (multiple file extensions such as .fasta and .fa) and FastQ (a version of the FastA data format that can also include information on sequencing) are the most popular text-based data formats for representing nucleotides and protein sequences. As text-based files, which in addition have a pre-ordered structure, they are machine-readable and a whole set of software tools in multiple programming languages are available to read and manipulate them.

In addition to FastA and FastQ data formats, SAM (Sequence Alignment Map), BAM (Binary Alignment Map), and CRAM (Compressed Reference-oriented Alignment Map) are all machine-readable data formats used for reporting sequence alignment in a standardised way. SAM is both human-readable and machine-readable, while BAM is a binary version of SAM and it is only machine-readable, but offers lossless compression of SAM files. CRAM files are similar to BAM files, but the compression algorithm used can range from lossless to lossy compression and therefore the files can be much smaller than the equivalent BAM files. Another data format worth mentioning in relation to sequencing data is the Variant Call Format (VCF), another type of text file format with a precise structure, and therefore machine-readable, that is specifically used to store gene sequence variations.

5.4 Images

As already mentioned in Section 3, images, with the exception of image-based data interfaces like QR codes, are unstructured data and therefore not machine-readable. Yet, image data are at the core of multiple human activities, such as medical diagnostics or material science, and improving image quality, processing image data, and analysing the information contained in them has been an ongoing effort for decades in the fields of digital image processing (Gonzalez & Woods, 2018) and computer vision (Szeliski, 2022). As a result, algorithms and workflows have been devised to extract at least some information from digital images. The examples that follows do not want to be comprehensive, but just to suggest a few of the possible methodologies currently available.

Before discussing them, however, a brief mention of what digital images are as computational objects is helpful. The image seen on a computer screen is stored in the computer memory as an array (a matrix in mathematical terms) of real numbers. Each element of the array represents a "picture element", a quantity better known as a **pixel**. You can easily view the individual pixels that form a digital image by zooming in. Each pixel is displayed as a coloured square (or a coloured rectangle more often on digital televisions). The colour of the pixel is set by the value(s) in the array. For instance, in a **grayscale image**, i.e., an image where the pixel only carries information on the light intensity and the visualised image only has black, white, and gray shades, each pixel can be represented using 8 bits (2^8 = 256) as an integer in the range [0, 255]. An **image in the RGB (Red-Green-Blue) colour representation**, instead, can be represented using 24 bits (8 bit for each colour channel) and giving for each pixel three values corresponding to the intensities in the Red, Green, and Blue channels. Image formats (e.g., JPEG, TIFF, BMP) differ in relation to compression standards (lossless, lossy). Lossy compression standards (e.g JPEG) can degrade image quality, but are smaller, a factor that can be very important for some use cases.

Once images are stored as arrays in the computer memory, they can be transformed using array operations and can be analysed to extract information from them. Gonzalez and Woods (2018) define a pragmatic scale of intervention on digital images, categorising **low-level processes** (operations that modify images, such as de-noising and contrast enhancement, and that result in a new transformed version of the original image), **medium-level processes** (operations that do not return a new image, but a set of machine-readable attributes extracted from the image, such as edges in a segmentation process), and **high-level processes** (operations that range from analysing extracted objects to prediction using AI models).

In the processing and analysis of images in biodiversity science, all the operations described above come into use. Low-level processes like image de-noising are helpful to improve the quality of wildlife images, mid-level processes such as image segmentation are the starting point for many workflows used in the mass digitisation of natural history collections to extract the various components of the

image (e.g., specimen, label, barcode, colour scale, etc.), AI models that recognise handwritten and typed labels are used to acquire the label content without manual intervention. Therefore, even as unstructured data, images remain a precious source of scientific information that can be extracted using automated and semi-automated strategies.

In conclusion of this section, it is important to add a comment related to digital images. Every equipment producing digital images - be it a camera, a scanner, a microscope, etc. -, generates alongside the image a list of metadata related to the image. These metadata can often be customised by the user and usually include a unique identifier for the image, its format specifications, information about the image creator and the creation date, data on the equipment and settings used to generate the image, image rights, etc. These pieces of information are crucial in the further reuse of the image data generated and should be always kept associated to the image (or at least in a database that can be easily consulted by the image user) and made available in one of the machine-readable formats discussed above (e.g., JSON, XML, etc.).

5.5 Other media

In biodiversity science, not only images, but also other media are routinely used for scientific research. Video data, for instance, can be employed to detect animals in ecology studies (Weinstein, 2018), audio data can be used for non-invasive biodiversity monitoring (Kvsn et al., 2020), and 3D-models of museum specimens can provide relevant information for studies of genetic and ecological bases of phenotypes (Blackburn et al., 2024).

These media data are distributed in a variety of proprietary and open data formats that differ considerably in relation to file size and quality, and hardware and software compatibility. For instance, popular formats for multimedia files based on ISO standards are MP4 (video) (ISO/IEC 14496-14:2020(en)) and MP3 (audio) (ISO/IEC 13818-3:1998(en). DICOM, the reference standard for medical imaging, is often used for volumetric 3D models.

Similarly to images, the other media data discussed in this section are not machine-readable and information can be extracted only by processing them. Machine learning models are especially popular to work with these data. For instance, several publications appeared in recent years discussing how YOLO **[4]**, which is a state-of-the-art model for object detection in videos, can be used for the needs of biodiversity science. Ad hoc models for object/sound detection (Müller et al., 2023) are also implemented, but often with mixed results due to lack of adequate training data. As already noted above, the devices that produce digital media also generate metadata and, at least these metadata, should be made available in a machine-readable format and provide all the information required for the use and re-use of the multimedia data.

[4]
YOLO (You Look Only Once) is a very popular family of models for object detection originally presented in 2015.

6. Formats for machine-readable data specific to biodiversity science

The biodiversity informatics community has invested considerable time and effort in the development of data standards specific to the needs of biodiversity science. The standards webpage curated by TDWG (Biodiversity Information Standards) is probably the best starting point to explore the data standards used by the biodiversity community.

As discussed in the Glossary, a data standard is typically associated to a data format, the element of interest in this guide. All data standards used in biodiversity science are associated to machine-readable data formats because for decades the biodiversity community has invested considerable time and effort in promoting tools that facilitate automatic data exchange and integration at scale. In this guide only a few popular standards used for biodiversity data exchange are explicitly discussed. For all other use cases endorsed by TDWG and not examined here please refer to the Standards section on the TDWG website.

- **DwC-A**

  Darwin Core Archive (DwC-A) is the data format associated to the Darwin Core standard. A Darwin Core Archive is a package of text files in CSV or TSV format stored in a zip folder. The data files are supplemented by an XML metafile (meta.xml) that provides information on the content and formatting of the data files in the archive and by a machine-readable metadata file compiled using a known standard, such as the Ecological Metadata Standard (EML).
  Each archive includes a core data file that stores all the data using a Darwin Core-compliant terminology and assigns a unique identifier to each data record. The identifier is used to link the information in the core file to the information available in the extension data files. DwC-A is the preferred format for data publication in GBIF and, as such, experience an overall high popularity in the biodiversity community.

- **ABCD and extensions**

  Alongside the DwC-A data format, GBIF also accepts data submissions that follow the specifications of the ABCD (Access to Biological Collections Data) standard and its extensions **[5]**. The ABCD standard is also maintained by TDWG and it is mainly conceived for exchanging

data about living and preserved natural history and botanical collection specimens and field observations.

For the ABCD data standard there is no data format analog to DwC-A. Data that follows the specifications of the standard, however, are always transmitted in XML, therefore they are in a machine-readable format. The structure of the XML data files follows the specifications and the terminology set out by the ABCD data standard.

[5]
The main extensions of the ABCD standard are ABCD EFG for the geosciences and ABCD DNA for genomic data.

- **Other machine-readable formats used in biodiversity science**

  Alongside the data formats mentioned above, which are based on TDWG data standards and regularly used for data exchange within the biodiversity community, there are other science-specific machine readable data formats that find application in biodiversity science. One example is the Network Common Data Form (NetCDF) that is both a data standard and a data format and has been specifically developed to deal with array-oriented scientific data.

  In biodiversity science, the NetCDF data format is used to share datasets on Essential Biodiversity Variables. Data in NetCDF format are stored in a single file structured as an header section followed by a data section. Building NetCDF files requires the use of the HDF5 library.

7. Machine-readable data in the age of AI: Some final remarks

The concept of AI-ready data is gaining increasing popularity in business as well as in science. The concept has no established definition and its reach ranges from practical considerations (e.g., data format, accessibility), to ethical considerations (e.g., lack of bias, accuracy). Regardless of how an organisation defines its AI-ready data, data machine-readability is a cornerstone of it. Only digital data that can be managed via automatic pipelines and consumed directly by a machine learning model or a deep neural network can be considered AI-ready. Making available data in a machine-readable format, therefore, is the first step to make them AI-ready.
Yet, if machine-readable data are easily consumed by an AI algorithm, this poses the question of what AI uses are considered lawful/acceptable for the data and how organisations can avoid a misuse of their data. Only proper data licensing, which also means making the data license machine-

readable, can offer a legal and practical protection in this case. All organisations, in and beyond biodiversity science, should devote special thought to license their machine-readable datasets appropriately.

8. References

Avram, E.D. (2003) Machine-Readable Cataloging (MARC) Program. In *Encyclopedia of Library and Information Science* 3, pp. 1712-1730. DOI: 10.1081/E-ELIS 120008993.

Blackburn, D.C. et al. (2024) Increasing the Impact of Vertebrate Scientific Collections through 3D Imaging: The openVertebrate (oVert) Thematic Collections Network, *BioScience* 74(3), 169-186. DOI: 10.1093/biosci/biad120.

Gonzalez, R.C. and Woods, R.E. (2018) Digital Image Processing (4th edition). Pearson Education, New York.

ISO/IEC STANDARD 29500-1: 2016(E) (2016) Information Technology - Document Description and Processing Languages: Office Open XML File Formats. Vernier, Geneva, Switzerland: ISO Copyright Office (available online).

Kvsn, R.R. et al. (2020) Bioacoustics Data Analysis – A Taxonomy, Survey and Open Challenges, *IEEE Access* 8, 57684-57708. DOI: 10.1109/ACCESS.2020.2978547.

Müller, J. et al. (2023) Soundscapes and Deep Learning Enable Tracking Biodiversity Recovery in Tropical Forests, *Nature Communications*, 14, 6191. DOI:10.1038/s41467-023-41693-w.

Sharma, G. (2016) Image-based Data Interfaces Revisited: Barcodes and Watermarks for the Mobile and Digital Worlds, *8th International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, India, pp. 1-6. DOI: 10.1109/COMSNETS.2016.74400213.

Szeliski, R. (2022) Computer Vision: Algorithms and Applications (2nd edition). Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-34372-9

Theissinger, K. et al. (2023) How Genomics Can Help Biodiversity Conservation, *Trends in Genetics* 39(7), 545-559. DOI: 10.1016/j.tig.2023.01.005.

OPEN Government Data Act H.R.1770 115th Congress (2017-2018) (available online).

Weinstein, B.G. (2018) Scene-specific Convolutional Neural Networks for Video-based Biodiversity Detection, *Methods in Ecology and Evolution* 9, 1435-1441. DOI:10.1111/2041-210X.13011.

9. Acknowledgements

This research was supported by the German Research Foundation (DFG) under grant number 528674292. I wish to thank Christian Bölling for his suggestions about machine-readable data formats specific to biodiversity science.