First steps towards the development of an integrated metadata management system for biodiversity-related micro-CT datasets

S. Faulwetter¹, N. Minadakis², K. Keklikoglou¹, M. Doerr², C. Arvanitidis¹

¹ Hellenic Centre for Marine Research, Institute of Marine Biology, Biotechnology and Aquaculture, Thalassocosmos, Gournes, 71500 Heraklion, Greece.
² Institute of Computer Science (ICS), Foundation for Research and Technology - Hellas (FORTH), Science and Technology Park of Crete, Vassilika Vouton, 71103, Heraklion, Greece

Aims

In an increasingly computerised world with distributed and interlinked information, metadata (i.e. data about data) have become of extreme importance. Without metadata, datasets are neither retrievable nor interpretable, and thus, the creation and management of metadata to describe a dataset is an essential tool for developing digital information (Hiom, 2004). During the last years, micro-CT has seen a strong increase of usage in biological research fields such as taxonomy, evolutionary and developmental research and functional morphology (e.g. Stoev et al. 2013, Wilhelm et al. 2011, Zimmermann et al. 2011). Hundreds, if not thousands of three-dimensional datasets of biological specimens have already been produced (Ziegler et al. 2011, Boistel et al. 2011). Through an ever-increasing effort of natural history museums to digitise their collections, this number is expected to be multiplied by several orders of magnitude over the next decades. However, the existence of most of these datasets remains unknown to the research community, since they are neither accessible through a public repository for micro-CT data, nor are they documented with metadata which could be queried (Ziegler et al. 2011, Boistel et al. 2011, Faulwetter et al. 2013). The development of standards and protocols for annotating and disseminating three-dimensional data as well as the creation of centralised registers to make the information retrievable remains an utmost priority for the micro-CT community working with biological specimens (Rowe and Frank 2011).

Standards for (micro)-CT data do exist – the most commonly used being the DICOM standard (Digital Imaging and Communications in Medicine, http://dicom.nema.org), which contains both format definitions and a communication protocol for the description and exchange of volume data. However, this standard is targeted to the medical sciences, and its integration with existing standards in the field of biodiversity information management is not straightforward. Biodiversity science has a need not only for documenting the actual scan data but also to link them to a species. Only through the species name can data be integrated into global species information systems such as the Encyclopedia of Life (http://www.eol.org), or combined with information on other sources, such as taxonomic species descriptions, information on the distribution of the species or genetic information. Several standards (e.g. developed by the Genomic Standards Consortium (http://gensc.org) for molecular data, or DarwinCore for specimen and observation data (http://rs.tdwg.org/dwc)) are already in widespread use in the respective communities, but they are not suitable for the annotation of three-dimensional morphological data. Stoev et al. (2013) used a combination of DarwinCore and the ISA-TAB format (Sansone et al. 2012) to describe both the scanning process and the specimen itself. This linked their micro-CT dataset through the species name to the wider biodiversity data network, but this approach arose from a lack of alternatives and was not an optimal solution.

In this paper, we present the first developments towards an integrated metadata management system for micro-CT data, with a focus on biological specimens. This system is being developed within the LifeWatchGreece infrastructure (http://www.lifewatchgreece.eu), a large-
scale project which aims at linking and integrating all data on the biodiversity of Greece—from
distribution data of species to literature resources, genetic information and three-dimensional
morphological information (micro-CT data). This placed the following requirements on the
metadata management system to be developed:

- Annotation through a standard which can be integrated with information from other
  sources (e.g. information on the biological classification of the specimen).
- Easy to use and not overly time consuming for researcher or technician
- Based on non-commercial / open source software
- Support for custom information on the specimen and its pre-scanning preparation
- Support for distributed storage of the actual datasets (can be on any server)
- Integrated, multi-faceted search across the metadata of all datasets, independent of the
  location of the dataset
- Web-based access to the metadata and data
- Compatible with Linux-based operating systems

Since no available software we were aware of fulfilled these requirements to our satisfaction,
we developed a customised solution. A standardised workflow for specimen handling,
metadata annotation and data management of micro-CT datasets was created and
subsequently translated into a digital information model, which allows the data to be integrated
into the global network of biodiversity-related information. Although parts of the software,
especially the user interfaces, are still under development and the LifeWatchGreece data
portal is not yet available to the public, we here present the first steps towards the
development of an integrated metadata management system for micro-CT scans.

Methods
The creation of the metadata management system involved the following major steps, which
are outlined in detail below:

A. Detailed analysis of the every-day workflow in the laboratory, identifying all actions and
   “objects” (persons, specimens, files) associated with each step of the work
B. Translation of the workflow into a digital model that defines entities and their relationships
C. Translation of the digital model into a formal ontology
D. Finalising software components, user interfaces and data flow

Of the above, steps A–C have already been completed, whereas step D is still under
development.

A. Laboratory workflow
The workflow associated with the creation of a 3D dataset is depicted in Figure 1 and
comprises the following steps (numbers correspond to encircled numbers in the figure):

1. Selection of the specimen and determination of its properties (e.g. specimen code, owner,
taxonomic identity, storage medium, etc)
2. Assignment of a unique specimen ID following a standardised format (see Figure 2)
3. Optional: Preparation of the specimen (e.g. if contrast enhancement is required)
4. Selection of the scanning container (e.g. SkyScan plastic holder, pipette tip, styrofoam)
   and scanning medium (air, water, ethanol)
5. Assignment of a unique scanning ID following a standardised format (see Figure 2)
6. Selection of the storage location for the files and creation of a standardised directory
   structure (based on the unique scanning ID)
7. Choice of parameters and scanning process
8. Choice of reconstruction parameters and reconstruction process
9. Storage of the specimen in a biological collection, to be kept as a physical reference for the scan
10. Optional: Creation of derived products (e.g. subvolumes, surface models, analyses, images, videos)
11. Final archiving of the files or deletion of the scan and documentation of the storage location (e.g. on server AB, on hard disk XY).

Figure 1: Schema of the standardised laboratory workflow. Numbers in circles are explained in the text.

B. Translation of the workflow into a digital model that defines entities and their relationships
The pieces of information resulting from the above steps can be grouped into four main categories: information on a) the specimen; b) preparation and scanning, c) reconstruction; d) subsets of data. These four groups relate to each other as follows: a specimen can be prepared and scanned one or multiple times, each of these scans can be reconstructed multiple times, and from each reconstruction multiple subsets can be created (Figure 2). Every specimen, scan, reconstruction and data subset is assigned a unique identifier, following a standardised format (Figure 2). These identifiers are used to link between the categories and thus connect e.g. a specimen to a specific reconstruction. In addition, we created a standardised filename and directory structure based on these identifiers to store projection files, reconstructions and derived data products (images, videos, surface models, analyses). For each of the four categories (or sets of actions) a set of additional metadata terms was defined, holding information on e.g. the identity of the specimen, its size, details on specimen...
preparation and scanning parameters, on persons performing the scan, on the purpose of scan and reconstruction, etc. (Table 1).

Figure 2: Relationship of the entities involved in or created by a scanning process. The format of the unique identifier is given in purple below the images. Details can be found in Table 1.

Table 1: Example of the metadata terms used to describe the scanned object, the scanning process and its products. In total, 57 metadata terms were defined.

<table>
<thead>
<tr>
<th>Term</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpecimenID</td>
<td>A unique identifier for the specimen in the format: mCT-xxxxx (where x = incrementing number from 00001 to 99999, with preceding zeros)</td>
</tr>
<tr>
<td>SpecimenProvider</td>
<td>Person who provided the specimen</td>
</tr>
<tr>
<td>SpecimenDescription</td>
<td>A short verbatim description of the specimen, which allows to understand the nature of the specimen at a glance (e.g. “dry ant”, “juvenile fish”, “mouse head”, “otolith of sea bream”)</td>
</tr>
<tr>
<td>Size_mm</td>
<td>Maximum length of specimen</td>
</tr>
<tr>
<td>FixationType</td>
<td>Original fixation medium of the specimen (e.g. formalin, ethanol)</td>
</tr>
<tr>
<td>PreservationMedium</td>
<td>Current preservation medium of the specimen (e.g. formalin, ethanol, air)</td>
</tr>
<tr>
<td>ScanID</td>
<td>A unique code of the format scan-xxxxx (where x = incrementing</td>
</tr>
<tr>
<td>Parameter</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>ContrastEnhancement Method</td>
<td>optional: short name of chemical treatment used for contrast enhancement</td>
</tr>
<tr>
<td>SampleHolder</td>
<td>A description of the sample holder</td>
</tr>
<tr>
<td>ScanningMedium</td>
<td>The medium that surrounds the sample during the scan (e.g. air, ethanol)</td>
</tr>
<tr>
<td>ScannedPart</td>
<td>(Body) part of the specimen that has been scanned</td>
</tr>
<tr>
<td>ScannedBy</td>
<td>Person who performed the scan</td>
</tr>
<tr>
<td>ScanDate</td>
<td>Date of the begin of the scanning process in the format YYYY-MM-DD</td>
</tr>
<tr>
<td>ScanningDuration</td>
<td>The duration of the scan in the format HH:MM</td>
</tr>
<tr>
<td>Instrument</td>
<td>The instrument on which the scan was performed</td>
</tr>
<tr>
<td>Voltage_kV</td>
<td>The voltage in kilovolt (kV)</td>
</tr>
<tr>
<td>Current_uA</td>
<td>The current in µAmpere</td>
</tr>
<tr>
<td>Filter</td>
<td>Filter used</td>
</tr>
<tr>
<td>Zoom_um</td>
<td>Resolution of the scan in µm (Zoom level)</td>
</tr>
<tr>
<td>CameraResolution</td>
<td>Camera resolution settings</td>
</tr>
<tr>
<td>Averaging</td>
<td>Frame averaging value</td>
</tr>
<tr>
<td>RandomMovement</td>
<td>Random movement value</td>
</tr>
<tr>
<td>360°</td>
<td>360° or 180° rotation scan</td>
</tr>
<tr>
<td>ExposureTime_ms</td>
<td>Exposure time in milliseconds</td>
</tr>
<tr>
<td>ReconID</td>
<td>A unique identifier in the form scan-xxxxx_rec-yy (scan-xxxxx= scanID as described above, yy=incrementing number within that scanID between 01 and 99, with preceding zeros)</td>
</tr>
<tr>
<td>ReconScope</td>
<td>Reason for reconstruction (e.g. first reconstruction, reconstruction of specific part, reconstruction with specific parameters).</td>
</tr>
<tr>
<td>ReconBy</td>
<td>Person performing the reconstruction</td>
</tr>
<tr>
<td>ReconDate</td>
<td>Date of reconstruction in the format YYYY-MM-DD</td>
</tr>
<tr>
<td>ReconDuration</td>
<td>Duration of the reconstruction in the format HH:MM</td>
</tr>
<tr>
<td>SubvolumeID</td>
<td>optional: A unique identifier in the form scan-0000x_rec-0y_voi-0z</td>
</tr>
</tbody>
</table>
(scan-xxxxx= scanID as described above, yy=incrementing number within that scanID between 01 and 99, with preceding zeros, zz=incrementing number within that reconstruction, between 01 and 99, with preceding zeros)

SubvolumeScope  optional: Content of the subvolume, why has it been created

SubvolumeCreatedBy  optional: Person who created the subvolume

C. Translation of the digital model into a formal ontology
While the developments of step B are a sufficient basis for a successful in-house data management, the metadata cannot easily be shared or integrated into a wider data network. To achieve this, the metadata model needs to be translated into an ontology (or semantic model) - a formalised naming description of entities, their properties and interrelationships that exist for the particular field of interest, expressed in a dedicated computer language. Computers can interpret these according to specific rules and deduce new information by combining data, thus applying “intelligent” reasoning. The selection of appropriate semantic models to be used for the representation of the defined concepts is a crucial step of the process. For the needs of the micro-CT scanning domain it was essential to use a model that can be used to a) describe biodiversity related concepts and events; b) describe events (event centric); and c) encode metadata about the steps and methods of production of digitisation (scanning) products and synthetic digital representations such as 2D or 3D Models. For these reasons, elements from the following three models of the CRM family were selected:

a) MarineTLO: MarineTLO is an ontology which was initially developed to describe concepts included in marine data sources, but is generic enough to describe any biodiversity-related information. It provides the necessary properties to relate facts on observational data with the respective spatiotemporal context and systematic (taxonomic) knowledge. It can be extended to any level of detail on demand (Tzitzikas et al. 2013, 2014).

b) Cidoc CRM: he CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing historical events at a human scale including scientific records. CIDOC CRM follows an activity centric pattern which is coming to all provenance models.

c) CRM Digital: CRM is an ontology specifically designed to encode metadata about the steps and methods of production (“provenance”) of digitisation products and synthetic digital representations such as 2D, 3D or even animated models created by various technologies (Theodoridou et al. 2010). CRM Digital is fully compliant with other, widely used, provenance models such as PROV-O but it is more detailed since it completely includes the initial physical measurement processes and their parameters.

Classes and properties of the models that were mentioned above were used in order to describe each step, object and property of the micro-CT scanning events. An example of the representation of the whole scanning workflow is shown in Figure 3.
Figure 3: Representation of the ontology classes /concepts and their relationships which describe the acquisition of a micro-CT image stack and derivatives. Abbreviations before the names of events, classes and metaclasses are internal identifiers.

D. Finalising software components, user interfaces and data flow

From the creation of the metadata to their standardised representation through an ontology a number of steps had to be performed which are outlined below. Their enumeration corresponds to the numbers in Figure 4.

1. **Creation of a set of metadata terms as described above.** During the current development phase, since not all software components are in place yet, a simple local database based on these terms was used to capture actual metadata.

2. **Transformation of the metadata into a structured format.** This produced CSV (comma separated value) files, simple text files in which the database fields are separated through a delimiter.

3. **Creation of the semantic model (ontology),** as described above. The elements (entities/concepts) and their properties and interrelationships of the ontology were described in a specialised data format / computer language, RDF (Resource Description Framework) and stored in a so-called schema file. This file basically defines how individual elements are related to each other and which properties they can have.

4. **Schema mapping.** Every metadata terms of the initial set of metadata was mapped against the semantic model’s classes and properties. A special mapping language, x3ml, was used for the implementation of the mappings, which are stored in a mapping file.

5. **Transformation of the metadata files into a compatible format (rdf) according to the mappings.** The x3ml mapping files and the metadata csv files were provided as input to the x3ml engine which transformed the csv files into RDF format according to the semantic model schema.
6. Ingestion of schema and metadata files into a Triple Store. Together with the schema files, the metadata (now in RDF format) were ingested into a semantic database (Virtuoso Triple Store), which acts as a metadata repository.

The following two steps were already designed but are not yet fully functional:

7. The metadata information can be made publicly available through the internet and enriched with other data from the Linked Open Data (LOD) cloud.
8. Users can query the data through a graphical user interface, available through the web.
9. Instead of exporting CSV files and importing them manually into the metadata repository, graphical user interfaces will be provided to the researcher and the data will be directly stored in the metadata repository. The option of uploading CSV files will continue to be provided to allow any external researcher to submit their metadata to the repository.

Figure 4: Schema of the steps involved in creating the metadata management system. Numbers in circles are explained in the text.

Results and Conclusions
We have completed the major steps towards the development of an integrated metadata management system for micro-CT data. While the individual components are not yet fully developed and connected and the data flow is not yet automatised, the individual components have already been used successfully: up to now, almost 600 micro-CT datasets have been fully annotated with metadata and stored in a standardised directory structure, and the metadata for these scans have been mapped to a specifically created ontology and imported into an integrated biodiversity data repository. Two major advantages became already obvious at this stage: a noticeable improvement of the laboratory workflow, and the power of data integration through the use of a semantic model.
Improvement of daily laboratory work
The introduction of a standardised workflow noticeably improved every-day lab work and research. Since all datasets are fully described, users can quickly re-assess specimens, preparations, scan setup and main scanning parameters. Datasets can be quickly located and retrieved thanks to a standardised naming and structure of files and directories. Although the set of metadata terms is by far not exhaustive (much more detailed information could be recorded during a scanning process), we deliberately restricted the terms to a “practical minimum” to reduce the time spent on the documentation process. With the current set of metadata, the documentation process requires minor additional effort by the researcher, yet the recorded information covers all important information on a scan that is needed to understand all processes and to interpret a dataset.

Advantages of the semantic model and information integration
Ontologies facilitate information exchange and integration between heterogeneous and distributed sources by providing a common understanding of concepts and their interrelationships. In our case this has two main advantages: a) The ontology can be used by any other user to describe their data and submit them into the LifeWatchGreece repository. Since the model is flexible, new concepts/terms can at any time be defined if the need arises (e.g. growing user base with different needs). Information can be omitted without breaking the model. This is an ideal basis for the creation of a public inventory of biodiversity-related micro-CT datasets, which is clearly missing from the biodiversity informatics landscape; b) The model enables semantic interoperability: since concepts are clearly defined and comparable, computers can deduce new knowledge by re-combining data from different sources, following specific rules. Here, it allows our micro-CT data to be integrated into a global network of linked information (the so-called Linked Open Data (LOD) cloud) and enriched with biodiversity information from other sources.

As an example for the possibilities of such information enrichment consider the following scenario: A user wants to query the system for all datasets showing specimens of the mussel family Mytilidae. However, the our initial metadata schema does not provide any terms for recording the biological classification of a specimen, apart from the species name. To retrieve the desired information, the user would first need to consult an external information source (e.g. Wikipedia, scientific literature) to obtain a list of all species in the family Mytilidae, and then query the metadata for each of these species names separately. If the information on the biological classification is, however, provided in another semantic repository, a computer can intelligently make the connection between these two different pieces of information and return the desired records by using information which was not explicitly mentioned in the original metadata (Figure 5). This is a powerful extension of the query possibilities for the data in our system.
Current and future developments
While the representation of data through a semantic model is highly beneficial for data standardisation and integration, the daily tasks of data entry, querying and data modifications require dedicated technical skills. Therefore, graphical user interfaces (GUIs) have to be developed that allow the user to interact with the data without the need for any such skills. The development of web-based GUIs that allow a) querying of the data, including enhanced information from the LOD cloud and b) to directly enter the data into the metadata repository (without the need for local storage systems and data transfer via CSV) is therefore one of the most important next steps. Furthermore, the web-based user interfaces will need to allow the download of actual scanning (image) data. Therefore, a data storage and archival system needs to be developed on the LifeWatchGreece servers. As the repository grows, thousands of scans will need to be stored and managed in the future. This requires a mechanism that ensures that metadata are always linked to the actual dataset, that they can be retrieved over the web and that datasets do not get lost and regular backups are created. It is expected that these developments will be finalised within the next year.

Acknowledgements
This study forms part of the ESFRI project LifeWatchGreece (MIS 384676) and was supported by the EU FP7 project MARBIGEN (FP7- REGPOT-2010-1).
References