DATA MANAGEMENT PLAN – FORM FOR GDAŃSK UNIVERSITY OF TECHNOLOGY EMPLOYEES

1. Data description and collection or re-use of existing data

1.1. How will new data be collected or produced and/or how will existing data be re-used?

To determine the phylogenetic relationships within the genus *Oncidium s.lato* several types of data will be collected:

- DNA sequences from the NCBI Nucleotide database (GeneBank) using BLAST web tool. All sequences deposited in this database can be downloaded and used for any purposes without any restrictions;
- DNA sequences of the genera samples obtained using DNA isolation from frozen leaves, polymerase chain reaction (PCR) and sequencing.
- Matrix with DNA sequences which will be compilation of data listed above (tools used to create it: FinchTV, AutoAssembler, ClustalX)
- Phylogenetic trees generated using PAUP tool and saved as vector images
- Additional laboratory notes, lists of samples, NCBI descriptions table

1.2. What data (for example the kinds, formats, and volumes) will be collected or produced?

There will be a few different types of files :

- DNA sequence data text files in FASTA format (.fasta)
- Data matrix saved as text file (txt)
- Images of the phylogenetic trees (.cdr)
- Lists, notes and other documents saved as standard office files, such as .docx and .xlsx.

All data won't exceed 1 GB of disc space.

2. Documentation and data quality

2.1. What metadata and documentation (for example methodology or data collection and way of organising data) will accompany data?

DNA sequences downloaded from NCBI database will be saved to the folders with names corresponding with the names of *Oncidium* genera. The name of every file will contain NCBI reference number and the access date. Accordingly to the FASTA format identifiers will be also part of sequences inside the files. Analogical structure of folders will be created for the sequences of analyzed samples. Laboratory research activities will be recorded using a selected laboratory diary tool and kept separately. Additional folders will be created for the DNA matrix and phylogenetic trees images. Every tree file will contain metadata describing the

parameters of the analysis and the bootstrap test.

All DNA sequences obtained during the samples' analysis will be uploaded to NCBI GeneBank and described accordingly to the FASTA format.

Rasterized tree images and other significant data will be deposited in the MOST Wiedzy Open Research Data Catalog – repository provided by the Gdańsk University of Technology. Additionally to the metadata description compatible with general metadata standards and stored in JSON-LD format, every dataset will contain 'readme' file containing information necessary to repeat the analysis.

2.2. What data quality control measures will be used?

All laboratory protocols will be based on methods presented in the scientific literature, manuals provided by the equipment and reagents producers, and good research practice. The analysis of DNA sequences will be performed by experienced researchers using specialized software.

To ensure the integrity, visibility, and general quality of shared research output different repositories were chosen for different types of data. Metadata records from both repositories are also indexed in the Data Citation Index (part of Web of Science platform).

3. Storage and backup during the research process

3.1. How will data and metadata be stored and backed up during the research process?

GeneBank sequences will be saved on the PI's computer. Laboratory data will be saved on the computers connected with the laboratory equipment and then transferred to the PI computer using USB flash drive. There will be a separated partition on the PI's laptop to store data associated with the project. To ensure backup the partition will be synchronized with the OneDrive (the enterprise version of the service ensures archiving data on the servers allocated within UE borders) and also the whole partition will be copied to the external hard drive once a week.

3.2. How will data security and protection of sensitive data be taken care of during the research?

No sensitive data will be collected during the research. Only authorized research team members will have access to the data and backup copies. All data saved on the laboratory computers accessible for all employees will be erased right after saving a copy on the PI's computer. Additionaly all drives will be protected by the password changed regularly.

4. Legal requirements, codes of conduct

4.1. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

4.2. How will other legal issues, such as intelectual property rights and ownership, be managed? What legislation is applicable?

The ownership and management of any intellectual property relating to the Project remain in the rights of the Gdansk University of Technology and the research team members accordingly to the Polish law and institutional regulations.

5. Data sharing and long-term preservation

5.1. How and when will data be shared ? Are there possible restrictions to data sharing or embargo reasons?

All DNA sequences obtained during the samples' analysis will be uploaded to NCBI GeneBank and described accordingly to the FASTA format. Rasterized tree images and other significant data will be deposited in the MOST Wiedzy Open Research Data Catalog – repository provided by the Gdańsk University of Technology. Data will be shared no later than the publication of the articles based on these data.

The way of data publication may be also associated with the different publishers' requirements but can't break the rules of open licensing and the date of datasets publication.

Due to project requirements and with the Gdańsk Tech authorities consent the data and results will be published in the open-access model under one of the Creative Commons licenses (CC0 where possible). Metadata will be always available without any restrictions (CC0). No embargo or any other restrictions are necessary.

5.2. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?

Due to the small disk space required for data storage, there is no need to select the data for long term archiving. All data, sequences and trees, but also notes or incorrect data, will be stored for at least 10 years after the project is finished and access to them will be possible only with the PI consent.

Most of the research output will be also accessible through the open research data repositories and there is no end date of sharing these resources.

5.3. What methods or software tools will be needed to access and use the data?

There is no need to gain specialized software to open and process data. The software for FASTA sequences files is provided by NCBI, tree images will be deposited in the one of open formats.

5.4. How will the application of a unique and persistent identifier (such us a Digital Object Identifier (DOI)) to each data set be ensured?

The datasets provided in the MOST Wiedzy repository will have the DOI assigned, sequences deposited in GeneBank will have GeneID

6. Data management responsibilities and resources
6.1. Who (for example role, position, and institution) will be responsible for data mangement (i.e the data steward)? Open Science Competence Center (pg.edu.pl/openscience) - established by GUT will be responsible for DMP and the quality of data deposited in the MOST Wiedzy repository. Project PI will be responsible for the procedures assessment and overall data quality.
6.2. What resources (for example financial and time) will be dedicated to data management and ensuring the data will be FAIR (Findable, Accessible, Interoperable, Re-usable)? There no need additional resources to ensure quality of data and compatibility with FAIR rules.