Spectral Image Processing with Efficiently Compressed TensoRs and AI

A Data Management Plan created using DMPonline.be

Creators: Ioannis Kalfas https://orcid.org/0000-0002-9957-1502, Wouter Saeys https://orcid.org/0000-0002-5849-4301, n.n. n.n.

Affiliation: KU Leuven (KUL)

Funder: Bijzonder Onderzoeksfonds

Template: KU Leuven BOF-IOF

Principal Investigator: Wouter Saeys https://orcid.org/0000-0002-5849-4301, n.n. n.n.

Data Manager: n.n. n.n., Ioannis Kalfas https://orcid.org/0000-0002-9957-1502

Project Administrator: n.n. n.n.

Grant number / URL: C2E/23/007

ID: 203523

Start date: 01-10-2023

End date: 01-10-2027

Project abstract:

Artificial Intelligence (AI), and especially convolutional neural networks (CNNs), has revolutionized the way images are processed thanks to their unique architecture involving many parameters, and the availability of massive amounts of labelled images for parameter estimation. Spectral imagers provide a fingerprint for each pixel, revealing important quality information that cannot be assessed with RGB cameras. This offers great added potential for quality control in agrofood. Despite their potential, spectral imagers also bring a high complexity and data dimensionality. In combination with the general unavailability of large, labelled datasets this requires strategic basic research into efficient learning strategies that allow to build, maintain and transfer efficient processing pipelines for analyzing these complex data structures with a limited number of labelled images.

Last modified: 14-03-2024

Spectral Image Processing with Efficiently Compressed TensoRs and AI

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		Indicate: ***N****(ew data) or* ***E****(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
Mustard_seeds		E	D	I	hdf5	<5TB	
Gluten_grains		E	D	I	hdf5	<1TB	
Disease_strawberries		E	D	I	hdf5	<1TB	
Fake_leaves		E	D	I	hdf5	<1GB	
3DXray_tomography		E	D	I	tiff	<5TB	
SpectralXray_CT		E	D	I	png	<5TB	
DL_models		N	D	M	pth	<100GB	
Xray_models		N	D	M	slx	<100GB	
CSV_files		N	D	N	csv	<1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

NA

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

All the new data stated on the data list will be potentially used as a basis for establishing prediction models for enhancing the way quality inspection is performed in the (Agro-)Food and Drink Industry (FDI). Hardware and software companies developing spectral imagers, sorting technologies for the FDI, or high-end AI solutions could benefit from the all data generated in this project.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

During SPECTRAI, large spectral image datasets will be generated, along with reference labels for each object. Data cubes will come in following size and formats:

- SXCT data (PNG or JPG format), (2-5 MB each; 5 GB per stack). Total expected size ~ 1 TB;
- HSI data (HDF5 format), (FX10: 350MB each, ~70GB, FX17: 200MB each, ~40GB) per case. Total expected size ~1 TB;

Data processing toolboxes will be written to be compatible with Python (building on backends in C++, Matlab, etc.), complemented by Avizo file formats for SXCT. Presentation files and manuscripts (reports, conference papers and journal articles) will be produced using MS Office and LaTeX.

A dedicated storage share is available on which each PhD student stores data files, with corresponding metadata in a cloud-based data management system. For the large volume of original HSI and SXCT data during and minimally 5 years after the project, we will rent large volume storage of KU Leuven. Costs will be covered by the project consumables budget. For daily work, a separate folder share is provided for the researcher with automatic back-up. Work-related files of the researcher will be on a OneDrive cloud folder (and GitLab cloud for code) that is daily backed up to a KU Leuven file server, where it is shared with the promotors. After conclusion of the PhD, files are transferred to a KU Leuven archive. For the active data management during the project, we plan to use the new ManGO platform, based on iRODS (open-source software), provided by KU Leuven-ICTS.

Will a metadata standard be used to make it easier to find and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- Yes

Metadata will be generated following the standards used in 'ManGO', the new KU Leuven system for data management. Metadata schemas will be created in collaboration with project partners, to make data reusable and findable within the ManGO platform.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- ManGO
- OneDrive (KU Leuven)
- Shared network drive (J-drive)

The data are stored on OneDrive accounts of specific users involved with the project and on ManGO (iRODS system provided by the Research Data Management and ICTS teams of KU Leuven).

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

If needed, more storage is provided by KU Leuven for our ManGO upon request.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data is kept in the KU Leuven Onedrive account of the PhD candidates, with access restricted to the account holders and granted to the IT department only in specific situations. As for ManGO, only a select group of users are allowed to view or change the data, but they must first be added as members of the group by one of the ManGO managers before they can do so. This is done by sending the users an invitation to join the group.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

Storing data in personal OneDrive folders does not incur any additional expenses. However, for ManGO, there is a fee of €35 per terabyte annually. It's anticipated that approximately 15 terabytes (TB) will be generated throughout this project, amounting to an annual storage expense of €525. These costs will be funded by the project funds of the associated partners participating in the project.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

Where will these data be archived (stored and curated for the long-term)?

- Other (specify below)
- Large Volume Storage (longterm for large volumes)

The data will be stored on ManGO for the long-term since it offers the lowest cost (€ 35 / TB / year) among other solutions like K or LVS (both at € 100,86 / TB / year). If required, it can always be moved to be archived on LVS or K.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

For ManGO, there is a fee of €35 per terabyte annually. It's anticipated that approximately 15 terabytes (TB) will be generated throughout this project, amounting to an annual storage expense of €525 or a total of €5,250 for a retention period of 10 years. These costs will be funded by the project funds of the associated partners participating in the project.
These costs will also be covered by the research groups involved.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- Yes, as embargoed data (temporary restriction)

It is our intention to make the publicly available for reuse some time after completion of the project, once they have been properly valorized.

If access is restricted, please specify who will be able to access the data and under what conditions.

During the embargo period, the data will only be accessible for members of the research groups involved in the project. Afterwards, the data will be made available to other researchers for non-commercial use.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

- Yes, intellectual property rights

The access may be restricted for some time (embargo) to protect the KU Leuven IP rights.

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other (specify below)

No decision has been made with respect to this.

When will the data be made available?

- Other (specify below)
- Upon publication of research results

Data that is used in potential publications will be available upon publication of research results.
Data that might be valorized will not be available until they are properly valorized.

Which data usage licenses are you going to provide?

If none, please explain why.

- CC-BY 4.0 (data)
- Data Transfer Agreement (restricted data)
- MIT licence (code)
- GNU GPL-3.0 (code)

Not decided yet. For code we consider both the MIT licence or the GNU GPL-3.0 license. During the embargo period, we will have to work with Data Transfer Agreements. Afterwards, we could move to CC-BY 4.0.

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

What are the expected costs for data sharing? How will these costs be covered?

Sharing data without incurring extra costs is expected as the data will be shared directly from their host locations. Expenses related to sharing will be assumed by the relevant research groups or the chosen repository, depending on the specific sharing arrangement selected.

Responsibilities

Who will manage data documentation and metadata during the research project?

The responsibility for managing data documentation and metadata falls upon the PhD researchers working on the project, with guidance from their respective promoters and data management representatives: Niels Wouters at MeBioS, Mariya Ishteva at NUMA, and Wannes Meert at DTAI.

Who will manage data storage and backup during the research project?

The responsibility for managing data storage and backup falls upon the PhD researchers working on the project, with guidance from their respective promoters and data management representatives: Niels Wouters at MeBioS, Mariya Ishteva at NUMA, and Wannes Meert at DTAI.
The backups of ManGO, the shared network storage and the settings of the OneDrive folder are managed by SET-IT and RDM teams of KU Leuven.

Who will manage data preservation and sharing?

The responsibility for managing data preservation and sharing falls upon the PhD researchers working on the project, with guidance from their respective promoters and data management representatives: Niels Wouters at MeBioS, Mariya Ishteva at NUMA, and Wannes Meert at DTAI.

Who will update and implement this DMP?

The responsibility for updating and implementing this DMP falls upon the data management representatives: Niels Wouters at MeBioS, Mariya Ishteva at NUMA, and Wannes Meert at DTAI.