BEYOND THE TREACHERY OF IMAGES: TRAINING AN ALGORITHM TO TURN ASTRONOMICAL OBSERVATIONS INTO PROBABILISTIC MODELS

Data Collection

What data will you collect or create?

In this project, we will create 20 TB of synthetic spectral line observations for a set of about 100.000 three-dimensional hydro-chemical models that were already created in another project (ATOMIUM).
The hydro-chemical models are stored as binary files, respectively of the hydrodynamics codes Phantom and MPI-AMRVAC.
To simplify the data manipulation all data files are first converted to HDF5 files and stored as such. Also the resulting synthetic observations will be stored as HDF5 files, which can easily be read, written, and manipulated by several programs, thus enabling us to easily share and reuse the data, also in the long-term.

How will the data be collected or created?

The data, i.e. the synthetic observations, will be created by applying the open-source radiative transfer model, called Magritte, to the existing hydro-chemical models.
Every model file has a unique name, based on the parameters of the model.
We will adopt the same structure for the corresponding synthetic observations.
The Magritte software library has an extensive test suite and was benchmarked against several other radiative transfer software libraries to ensure the quality of the results.
All our software pipelines use git for version control and are publicly available on GitHub.

Documentation and Metadata

What documentation and metadata will accompany the data?

The hydro-chemical models and their corresponding synthetic observations will be made publicly available through a website that explains their structure, and gives examples on how they can be used. This website will probably be hosted at KU Leuven, but this is still being discussed. The Magritte documentation (that we developed previously) will act as an example for this.
In their metadata, the synthetic observations contain the version number of Magritte that was used to create them, together with all the parameters of the hydro-chemical model corresponding to that synthetic observation.

Ethics and Legal Compliance

How will you manage any ethical issues?

Since we create all data ourselves with tools that we developed ourselves, we do not foresee any ethical issues.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

I, Frederik De Ceuster, as PI of this project will own the data, and will make it publicly available under the CC BY-SA license. Hence, no restrictions will be put on third-party reuse other that working under the same license.

Storage and Backup

How will the data be stored and backed up during the research?

All data is stored on the systems of the Institute of Astronomy at KU Leuven, which is regularly backed-up daily, both on site and off site.

The Institute of Astronomy has dedicated staff who are responsible for the maintenance of these systems.

How will you manage access and security?

Since everything will be publicly available from the start, we do not expect any security issues.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Since the data will be used to train radiative transfer algorithms, in principle all of our data can have value over the next 10 years.
As one of the objectives of our research will be to find efficient ways to represent the information stored in our data, or in other words find ways to compress the data, if successful, we might be able to store only the compressed version of our data.

What is the long-term preservation plan for the dataset?

We expect our data to be relevant for the next 10 years and will provide them on a self-hosted website. We budgeted 6000 EUR to buy the required storage and back-up for this (20 TB at 300 EUR/TB). This is funded and we are currently discussing how to best use this budget.

Data Sharing

How will you share the data?
The data will be made available on a self-hosted website which will be advertised in our publications and presentations.

Are any restrictions on data sharing required?
No restrictions on data sharing are needed.

Responsibilities and Resources

Who will be responsible for data management?
I, Frederik De Ceuster, the PI of this project, will be responsible for the data management.

What resources will you require to deliver your plan?
For the long-term storage of the data in this project, and in particular the self-hosted website that makes the data available on the web, additional storage capacity is necessary on the systems of the Institute of Astronomy. For this purpose, 6000 EUR was budgeted in the funding this project, which was granted and will be used once it is clear where and how we will host the website. The technical expertise is currently already available at the Institute of Astronomy to deploy this.

Planned Research Outputs

Software - "Probabilistic Deprojection Tool"

A software tool to turn spectral line images into probabilistic 3D models.

Dataset - "ATOMIUM Synthetic Observations"

Synthetic observations of the ATOMIUM hydro-chemical models.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Probabilistic Deprojection Tool	Software	2025-10-30	Open	GitHub		Creative Commons Attribution Share Alike 4.0 International	None specified	No	No
ATOMIUM Synthetic Observations	Dataset	2025-10-30	Open	None specified	20 TB	Creative Commons Attribution Share Alike 4.0 International	None specified	No	No