




DaedalusData: A Dockerized Platform for Exploration, Knowledge Externalization and Labeling of Image Collections

Alexander Wyss ^{1,2}

¹ Independent Researcher ² Roche pRED

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [George K. Thiruvathukal](#) 

Reviewers:

- [@tushardave26](#)

Submitted: 20 March 2025

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#))

Summary

DaedalusData is an open-source visual analytics platform designed for interactive exploration, knowledge externalization, and labeling of large image collections. It provides researchers and data scientists with a flexible and accessible environment for analyzing image datasets without requiring advanced technical skills. The system combines dimensionality reduction techniques for visual exploration with interactive filtering and selection tools, enabling users to label images efficiently and externalize their domain knowledge through customizable label alphabets. DaedalusData is fully dockerized for easy deployment and includes integrated Jupyter notebooks for feature extraction and dimensionality reduction, making it accessible to users without extensive programming or infrastructure expertise.

Statement of Need

The visual exploration and labeling of large image collections is a common challenge across diverse scientific fields. Researchers frequently need to analyze thousands of images, discover patterns, extract knowledge, and create labeled datasets for downstream analysis. While numerous tools exist for specific aspects of this workflow, there remains a gap. DaedalusData fills this gap for analysis tasks that are in a very explorative phase, where the user still needs to develop the label alphabets and explore the data to understand the patterns.

Existing image exploration tools often require significant technical expertise to install and operate, focus exclusively on either visual exploration or labeling (through active learning), or lack the flexibility needed for domain-specific adaptations. Commercial tools can be expensive and closed-source, limiting accessibility and customization. Meanwhile, custom implementations require considerable development effort and often result in solutions that aren't generalizable across disciplines.

DaedalusData addresses these challenges by providing a unified, accessible platform that enables researchers to:

1. Explore large image collections through interactive visualizations based on image content and metadata
2. Create and maintain multiple label alphabets that reflect diverse domain perspectives
3. Efficiently label images through interactive selection tools and similarity-preserving projections
4. Externalize expert knowledge through semi-supervised label-informed projections
5. Deploy the entire system with minimal setup via Docker, requiring no specialized infrastructure

By combining these capabilities in an easy-to-deploy package, DaedalusData democratizes

40 access to sophisticated image analysis techniques, enabling researchers across disciplines to
41 gain insights from their image collections more efficiently.

42 Architecture and Functionality

43 System Overview

44 DaedalusData consists of two main components:

- 45 1. **Frontend:** A Nuxt-based web application that provides the user interface for exploration
46 and labeling
- 47 2. **Jupyter Environment:** Integrated notebooks for feature extraction and dimensionality
48 reduction

49 The entire system is containerized using Docker, allowing for consistent deployment across
50 different environments. Users interact primarily through the web interface while having the
51 option to customize feature extraction and dimensionality reduction parameters through the
52 Jupyter notebooks.

53 Data Organization

54 DaedalusData operates on a simple, file-based data structure with mounted directories:

- 55 ■ **Images:** Original image files (PNG format)
- 56 ■ **Metadata:** JSON files containing attributes associated with each image
- 57 ■ **Features:** Extracted features in CSV or NPZ format
- 58 ■ **Projections:** Dimensionality reduction results for visualization
- 59 ■ **Labels:** User-defined label alphabets and assignments

60 This straightforward, file-based approach enhances interoperability with other tools in scientific
61 workflows and eliminates the need for database expertise or complex backend infrastructure.
62 Researchers can easily inspect, modify, or integrate their data with other scientific tools.

63 Key Features

64 Interactive Image Exploration

65 DaedalusData provides two primary exploration modes:

- 66 1. **Projection View:** Displays images in a 2D space based on dimensionality reduction of
67 selected attributes
- 68 2. **(Semi-)Supervised Projection View:** Include one or more label alphabets into the
69 dimensionality reduction process, allowing for label-informed projections that improve as
70 the user externalizes knowledge.

71 Both views support interactive zooming, panning, and filtering, allowing users to navigate
72 through thousands of images efficiently. Users can customize the visualization by adjusting
73 image size and transparency to reduce visual clutter.

74 Knowledge Externalization

75 A key innovation in DaedalusData is its support for knowledge externalization through label-
76 informed projections. As users assign labels to images, these labels can be incorporated as
77 additional attributes in the dimensionality reduction process, creating projections that reflect
78 both image features and expert knowledge. This creates a feedback loop where:

- 79 1. Initial projections guide users to discover patterns
- 80 2. Users label images based on discovered patterns
- 81 3. Label-informed projections reveal new patterns incorporating expert knowledge

82 4. Additional labels are created, further refining the projections
83 This iterative process enables the progressive enrichment of the dataset with expert knowledge.

84 **Efficient Labeling**

85 DaedalusData accelerates the labeling process through:

- 86 ■ Multi-selection tools for labeling many images simultaneously
- 87 ■ Label alphabets to organize related labels into meaningful collections
- 88 ■ Persistence of selections across different views and projections
- 89 ■ Visual encoding of labeled images for easy identification

90 The combination of efficient labeling tools with exploratory visualization significantly reduces
91 the time required to create high-quality labeled datasets.

92 **Integration with Scientific Workflows**

93 DaedalusData integrates with existing scientific workflows through:

- 94 ■ Template Jupyter notebooks for customizable feature extraction using standard libraries
95 (e.g., TensorFlow, scikit-learn)
- 96 ■ Mounted volumes for easy data interchange with other tools

97 **Usage Examples**

98 DaedalusData was originally designed for analysing single object images, but has been applied
99 to diverse image analysis tasks since, including:

- 100 1. Medical image analysis for identifying patterns in diagnostic images
- 101 2. Materials science for classifying microscopy images
- 102 3. General-purpose image exploration and labeling tasks

103 In each case, the system enabled researchers to interactively explore their image collections,
104 discover meaningful patterns, and efficiently create labeled datasets for further analysis.

105 **Conclusion**

106 DaedalusData fills an important gap in the scientific software ecosystem by providing an
107 accessible, comprehensive platform for the exploration, knowledge externalization, and labeling
108 of image collections. By combining interactive visualization, efficient labeling tools, and
109 easy deployment through Docker, DaedalusData enables researchers across disciplines to gain
110 insights from their image data more effectively.

111 **Acknowledgments**

112 This work builds upon research originally published in IEEE Transactions on Visualization and
113 Computer Graphics (Wyss et al., 2025).

114 **References**

- 115 (Wyss et al., 2025): Wyss, A., Morgenshtern, G., Hirsch-Hüsler, A., & Bernard, J. (2025).
116 DaedalusData: Exploration, Knowledge Externalization and Labeling of Particles in Medical
117 Manufacturing — A Design Study. IEEE Transactions on Visualization and Computer Graphics,
118 31(1), 54-64. <https://doi.org/10.1109/TVCG.2024.3456329>

119 Wyss, A., Morgenshtern, G., Hirsch-Hüsler, A., & Bernard, J. (2025). DaedalusData: Explo-
120 ration, knowledge externalization and labeling of particles in medical manufacturing — a
121 design study. *IEEE Transactions on Visualization and Computer Graphics*, 31(1), 54–64.
122 <https://doi.org/10.1109/TVCG.2024.3456329>

DRAFT