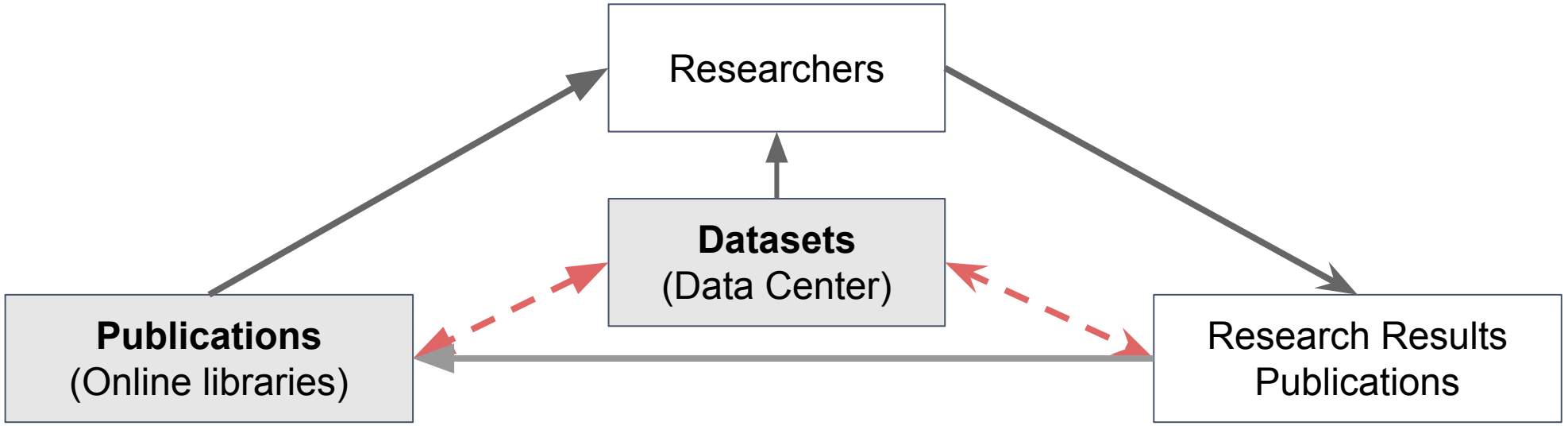
Application of ML/AI for Identifying Earth Science Datasets in Research Publications

Irina Gerasimov, Jacob Atkins, Edward Jahoda, Andrey Savtchenko, Jerome Alfred, Jennifer Wei

Goddard Earth Sciences Data and Information Services Center (GES DISC)
NASA Goddard Space Flight Center, Code 610.2

2nd NASA AI workshop, February 9-11, 2021

Connecting Datasets and Research (open science problem)



When scientific publication is not directly linked to the data it used the researchers and data producers face the problems such as:

- **Reproducibility** of the research results
- **Provenance** of created data
- **Attribution** of research results to used data
- **Findability** of datasets used in research

Connecting Datasets and Research (solution)

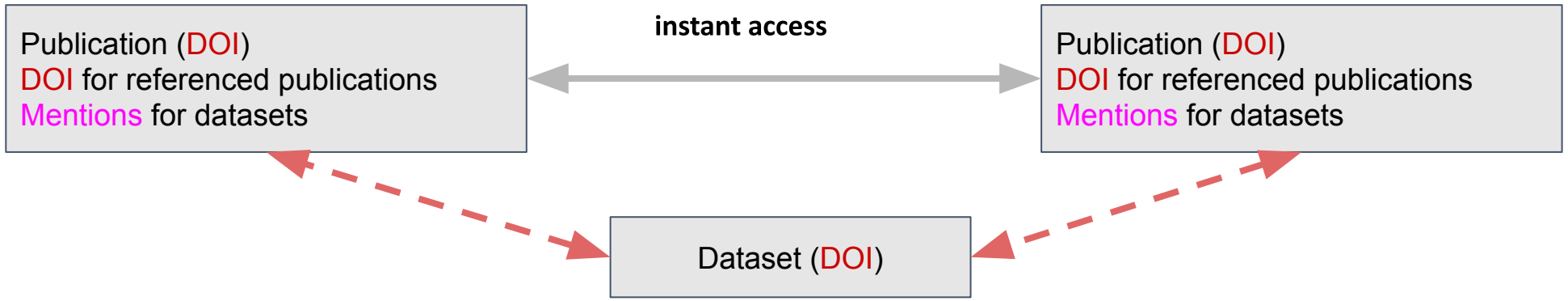
Solution: Developing a library of the citations that provides ***direct link between publications and datasets*** and/or missions/instruments, models and projects that produced the data archived at a data center.

Multiple benefits**:**

- Dataset science impact metrics.
- Credit to the dataset creators.
- Provenance: input and output datasets.
- Dataset usage-based discovery.
- Dataset recommendation.
- Dataset usage disciplines and topics.
- Dataset applications.

Connecting Datasets and Research (challenges)

- Digital Object Identifiers (DOIs) allow **instant** access to referenced publications or datasets.
- Using DOIs is a well established practice to reference other publications, however, authors do not follow this practice to reference datasets' DOIs:
  - *While NASA data centers introduced DOIs ~10 years ago, less than 20% of research papers published in 2019 provided DOIs for GES DISC datasets.*
- To reference datasets, publication authors use mentions rather then DOIs.



Dataset identification in publications

Currently - Manual: Subject matter experts (SMEs) read publications and identify the datasets mentioned there by various dataset attributes:

- Mission and instrument (for observational data) or model (for reanalysis data)
- Dataset processing level
- Spatial and temporal resolution of the dataset
- Dataset measurements and variables
- Name of the dataset creator

Goal - Automated:

Utilize attributes listed in the dataset metadata to extract terms from the publication texts and identify datasets through the means of ML/AI methods.

Publication's text preparation

1. Convert publication's PDF to ASCII (using Cermine)
2. Retain paper sections that most likely contain mentions of the datasets that were actually used in the paper
   a. **Eliminate: Abstract** -- *may* contain mission, instrument or model names, and names of some significant measurements and variables -- good for preliminary paper classification.
   b. **Eliminate: Introduction** -- mentions of datasets used in previous research
   c. **Retain: Main paper body** and **Acknowledgements** -- mentions of the datasets used in the paper
   d. **Eliminate**: **References** -- mentions of datasets used in previous research. May contain DOIs or full citations of the datasets used in the publication (then they can be easily extracted).

Term extraction using GCMD Ontology

NASA data archives use Global Change Master Directory (GCMD) ontology terms to create dataset metadata: a set of GCMD terms can identify single dataset or a group of datasets.

NASA GCMD ontology provides extensive coverage of names of missions, instruments, reanalysis models, and measurements.

Idea: employ GCMD ontology to extract terms from publication's text, e.g.:

Measurement Mission Instrument

This is consistent with the diurnal variations in ClO VMR observed by UARS MLS.

Based on the terms extracted from the sentence use heuristics to determine the dataset.

Results: low precision, high recall.

Named Entity Recognition

When manually examining the publications, SMEs determine the datasets used in the paper by finding the mentions that identify datasets, e.g. :
... *MLS Version 4, Level 2 CO data*… mention indicates the following dataset: "MLS/Aura Level 2 Carbon Monoxide (CO) Mixing Ratio"

Idea: record mentions identified by SMEs and train NER model to extract similar mentions from the text. Then use heuristics to determine dataset candidates.

Implementation: The NER model from the Allen Institute for AI's open source collection of NLP Models was trained with the mentions collected by SMEs.

Results: High precision, low recall.

Lessons learned: Ontology Labeling vs NER

	Ontology Labeling	Named Entity Recognition
Pros	**Fully automated** Works well for the datasets that can be differentiated by a small number of terms.	**High precision** -- it is more important to correctly identify dataset then miss one.
Cons	Terms have to be weighted in their ability to identify dataset. All terms extracted from the sentence are treated as “Bag of Words” - this can produce combinations for multiple datasets which may result in low precision.	Requires manual mentions collection. Each SME labels mentions differently which affects model precision. Does not work well for sentences with more than one mentioned dataset - low recall.

- In many cases even SMEs cannot determine exact dataset used in the publication.
- Information about mission/instrument, model or project the data were used from as well as key measurements are still very important.

Current work

- Continuing to improve term extraction with various NLP methods.
- Instead of mentions, identifying sentences that contain the most significant terms such as mission/instrument and model or project names.
- Extracting other terms from those sentences.
- Use all extracted terms as input into a predictive model classifier.
- Evaluating ML similarity measurements between identified sentences and dataset names.

Ultimate goal: Fully automated labeling pipeline

As hundreds of publications produced each year and added to GES DISC library, our goal is to create a fully automated pipeline that generates citations labels identifying the data used in that paper: