IN23D-0898 - Reusing Data and Metadata to Create New Metadata through Machine-learning & Other Programmatic Methods

Justin Gosses[1,2], *Anthony R. Buonomo*[2,3], *Brian A. Thomas*[4], *Evan Taylor Yates*[4,5] *and Rena W. Yuan*[6], *(1)Science Application International Corporation Houston, Houston, TX, United States, (2)NASA Headquarters, Washington, DC, United States, (3)Science Applications International Corporation Washington DC, Washington, DC, United States, (4)NASA Headquarters, Washington, United States, (5)Science Applications International Corporation, Mountain View, United States, (6)U.S. Department of Agriculture, Washington D.C., United States*

Abstract

Recent improvements in natural language processing (NLP) enable metadata to be created programmatically from reused original metadata or even the dataset itself. Transfer-learning applied to NLP has greatly improved performance and reduced training data requirements.

In this talk, we'll compare machine-generated metadata to human-generated metadata and discuss characteristics of metadata and data archives that affect suitability for machine-learning reuse of metadata.
Where as human-generated metadata is often populated once, populated from the perspective of data supplier, populated by many individuals with different words for the same thing, and limited in length, machine-generated metadata can be updated any number of times, generated from the perspective of any user, constrained to a standardized set of terms that can be evolved over time, and be any length required. Machine-learning generated metadata offers benefits but also additional needs in terms of version control, process transparency, human-computer interaction, and IT requirements.
As a successful example, we'll discuss how a dataset of abstracts and associated human-tagged keywords from a standardized list of several thousand keywords were used to create a machine-learning model that predicted keyword metadata for open-source code projects on code.nasa.gov. We'll also discuss a less successful example from data.nasa.gov to show how data archive architecture and characteristics of initial metadata can be strong controls on how easy it is to leverage programmatic methods to reuse metadata to create additional metadata.