Harvester-Curator, a tool to elevate metadata provision in data and/or software repositories

Sarbani Roy, Cluster of Excellence SimTech, University of Stuttgart, Germany

In today's scientific landscape, metadata plays a pivotal role by providing crucial context and structure to raw data, encompassing details such as origin, format, and provenance.Given the challenge of identifying the relevance of information in terms of metadata, manual procedures often dominate the compilation process, posing challenges for researchers due to time constraints, a lack of guidelines, and perceived complexity. Consequently, there is a compelling need for the development and implementation of an automated metadata collection process to simplify this intricate facet of research data management. This presentation introduces Harvester-Curator, a Python-based tool designed to elevate metadata provision in data and/or software repositories, representing a noteworthy advancement in metadata provisioning. This tool streamlines the transition from local folders to datasets within repositories.

The initial stage, termed "Harvester," operates as a scanner, navigating through user-specified folders, categorizing files based on types, and identifying appropriate parsers for diverse file formats.These parsers are tailored to extract metadata from various structured file formats, methodically retrieving and consolidating information into a well-organized JSON file. Operating locally, this phase can collect metadata from large files often omitted from version control, offering researchers a streamlined and automated approach to comprehensive metadata collection.

In the subsequent stage, Harvester-Curator evolves into a curator, utilizing gathered metadata to fill in metadata fields within a designated repository. The goal is to allocate harvested metadata to relevant blocks, minimizing metadata loss and ensuring proper arrangement within the appropriate parent field and metadata schema of the target repository. The tool also provides choices for plug-in mappings, establishing connections between the attributes of harvested metadata and the metadata fields of the target repository.

Harvester-Curator underwent thorough testing using demoDaRUS, a test server associated with DaRUS (Data Repository of the University of Stuttgart), showcasing its robust functionality. Ongoing efforts aim to expand its capabilities to include other Dataverse installations and well-known repositories such as Zenodo. The tool's inherent extensibility allows seamless integration of new parsers, rendering it versatile and adaptable to the evolving requirements of research. This tool not only streamlines the

process of metadata collection but also significantly contributes to enhancing data accessibility and interoperability within repositories.