Título da comunicação: RODA-in - a generic tool for the mass creation of submission information packages

Resumo:

The latest advancements in digital technology encouraged companies and institutions to use and create data in a digital format. In addition to the new digital data, a large portion of previously created information is being digitalized in order to facilitate its preservation, access and use on the business process that are also digitally managed. However, the digital tech introduces new issues in the storage of information, which can be related to hardware failures, obsolete technologies, loss of cultural and social context or economical politic reasons. Digital preservation was developed to counter this problematic, which consists in the capacity of guaranteeing that the digital information remains accessible and that it can be interpreted in the future using a different technological platform from the one used in the moment of its creation.

Although concepts in this area already matured, e.g. the concepts of submission information packages (SIP) and ingest processes are defined on the Reference Model for an Open Archival Information System (OAIS), there has been a lack of open source tools to support the Producers in the creation of such packages. One common approach is the development of specific integrations between Electronic Records Management Systems (ERMS) in use at the Producers side and the archive, allowing to create SIPs directly from these systems and submit them to the archive's ingest workflow. However, a considerable number of Producers do not use a document management system at all and simply manage their records on a shared folder on the local network. When systems integration is not possible due to lack of support, expertise or budget, the alternative is to create SIPs manually.

The approach taken was to create an offline application to empower producers and archivists to create SIPs ready to be submitted to an OAIS. The file system works as the source of information which is used in semi-automated SIP creation process based on rules and heuristics.

The workflow can be described in five easy steps:

1. Choose the folder that holds the data to be archived
2. Import or create a new classification scheme
3. Create a rule to transform the input data to SIPs
    a. Choose SIP-creation aggregation rule from the 4 existing options
    b. Choose the descriptive metadata association rule from the 4 existing options
4. Manually edit the created SIPs (Optional)
5. Export the SIPs in the desired format

There are additional steps that the user can take to tweak the output like preventing some files to be added to the SIPs and editing or adding new descriptive metadata. One of the goals of this tool is to provide a simple approach to enable users to create thousands of SIPs with just a few mouse clicks, but also to enable advanced features that satisfy the needs of more demanding users.

The application has a file explorer capable of handling large amounts of data efficiently, as well as hiding and showing items marked in different states. Classification plans (which are a hierarchical collection of classification levels) can be imported from a file or created directly in the application. After the user has created a classification plan it can be exported for later use.

SIPs are created using a rule which is defined by a set of files/folders, one aggregation rule and one metadata association rule. The files and folders are selected in the application's file explorer, which should then be dragged to the classification plan to start the creation of the rule. There are currently four aggregation options and four metadata association options in RODA-in which should be enough for most of the cases, but more can be added if needed. Optionally, the user can manually edit the SIPs using functionalities like: editing the existing metadata, adding new metadata files, adding/removing files from the SIP's content, adding documentation, etc. Finally, the SIPs can currently be exported in one of two formats: BagIt or E-ARK SIP, although more can be added in the future.

To conclude, the application is being piloted by the E-ARK project associations and the Portuguese National Archives, which have given positive feedback and confirmed the tool solves the mass package creation problem.

Nota biográfica:

André Diogo Ribeiro Assunção Pereira
KEEP Solutions
adrapereira@gmail.com

André Pereira is a finalist of the Masters in Informatics Engineering, University of Minho, with specializations in the areas of Language and Knowledge Processing, Computer Graphics and Distributed and Parallel Computing, and is writing his master's dissertation in the mass creation of information packages. Bachelor in Informatics Engineering by the same university, André is currently employed as an Analyst and Programmer at KEEP Solutions and is involved in the E-ARK project, a European project with research in the digital preservation field. André is currently responsible in the development of RODA-in.