Scientific data management: A life cycle view

Ibrahim Elsayed, Adnan Muslimovic and Peter Brezany
Institute of Scientific Computing, University of Vienna, Austria

Introduction

Scientific data are being collected to a great extent in various research domains. They are stored on multiple national sites in various data repositories. Scientific collaborations are targeting to provide access to these primary data by the means of an e-infrastructure. Through portals scientists are able to utilise these data for significant analyses in the context of their interest. The output of these analyses aims at defining a large number of predictions and might provoke further experimentation, which in turn may take days or weeks, depending on computational and human resources available. However, the resulting data – called derived data – that have arisen from the research task represents valuable information not only to the acting research group, but also to other groups with respect to other research areas.

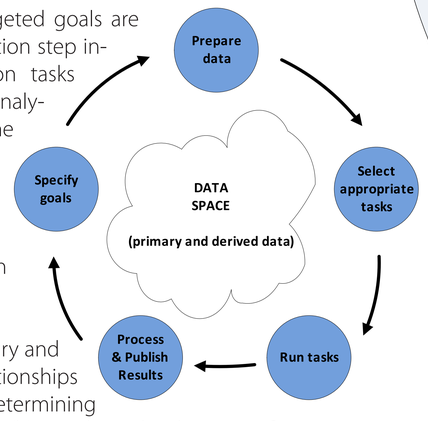
Objectives

Main objective is to link derived data with their corresponding primary data in e-Science applications by providing semantically rich relationships. Further, to make both relationships and data available within a space of data for people from various groups of organizations who might have use of it and who want to collaborate by the means of virtual organizations in the context of an e-infrastructure. To fulfill these goals, we further develop the dataspace paradigm introduced in [1,2].

The e-Science life cycle

The e-Science life cycle is defined as a domain independent ontology-based iterative metamodel, tracing semantics about procedures in e-Science applications. Iterations of the model - so called e-Science life cycles - organized as instances of the e-Science life cycle ontology, are feeding a dataspace, allowing the dataspace to evolve and grow into a valuable, intelligent, and semantically rich space of scientific data.

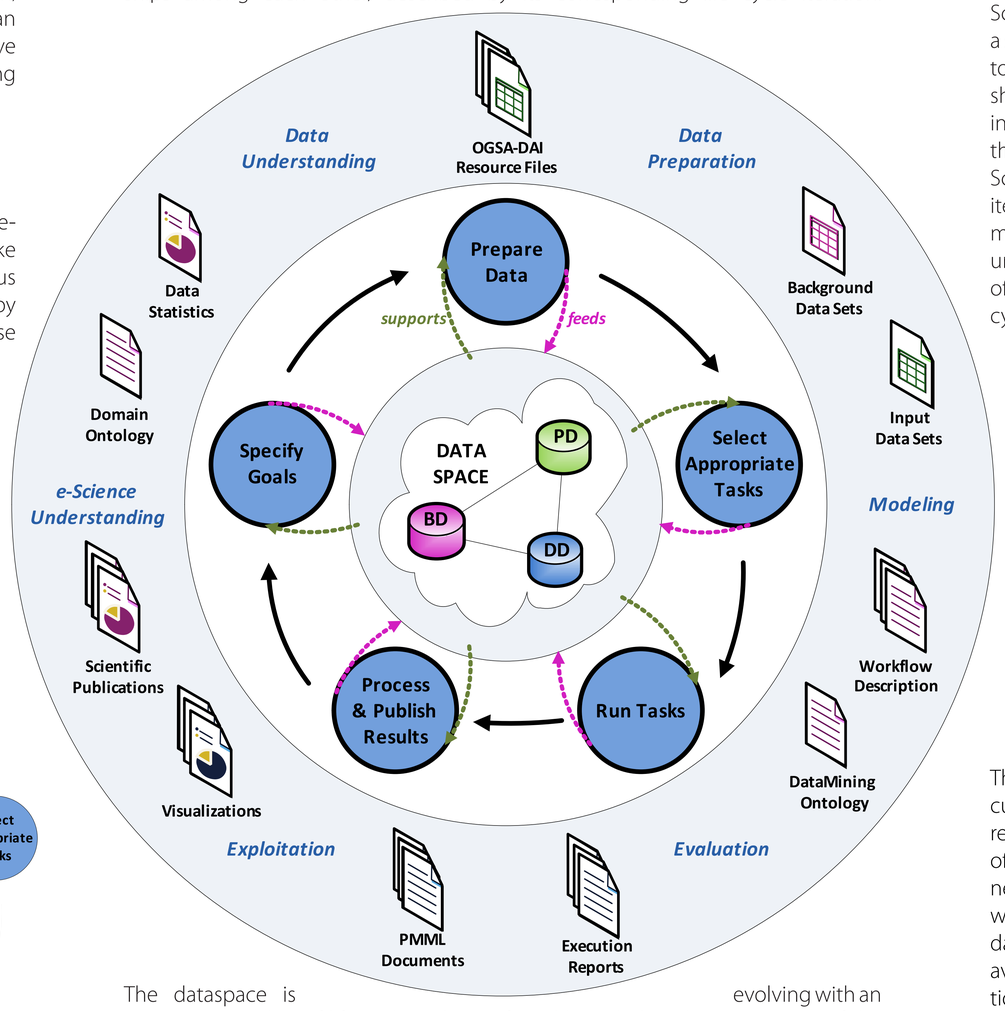
At the beginning of the life cycle targeted goals are specified, followed that a data preparation step including preprocessing and integration tasks is fulfilled. Further appropriate data analysis tasks are selected and applied on the prepared dataset of the previous step. Finally, achieved results are processed and published, which might provoke further experimentation and consequently specification of new goals within the next iteration of the life cycle.



The outcome of this is a space of primary and derived data with semantically rich relationships among each other providing (a) easy determining of what data exists and where it resides, (b) searching the dataspace for answers to specific questions, (c) discovering interesting new data sets and patterns, and (d) assisted and automated publishing of primary and derived data.

Methods

One iteration of the e-Science life cycle has, in short, a goal specification, a set of input data (primary data - PD), a set of output data (derived data - DD), a set of background data - BD, and a set of activities describing what has been done to the input data sets in order to produce the output data sets. These data sets are populating the scientific dataspace, enriched with semantic relationships among each other, described by its corresponding life cycle iteration.

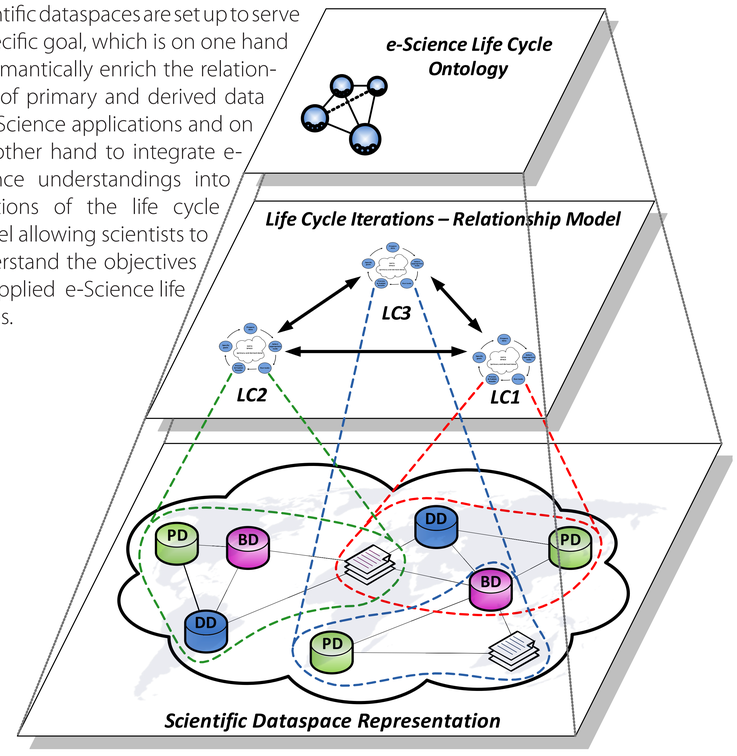


The dataspace is evolving with an increasing number of life cycles. The profound knowledge about iterations of the e-Science life cycle, consolidated within instances of the ontology represents an intelligent relationship model for scientific dataspaces, because it provides (a) creation, (b) representation, and (c) searching of semantically rich relationships among dataspace participants.

Results

At first, supported by the e-Science life cycle ontology, which organizes the concepts and coherences of the e-Science life cycle activities, a metamodel independent from the various e-Science domains is set up. Then this metamodel is applied to describe domain-specific iterations of the e-Science life cycle, which describe the relationship among data participating within the scientific dataspace.

Scientific dataspaces are set up to serve a specific goal, which is on one hand to semantically enrich the relationship of primary and derived data in e-Science applications and on the other hand to integrate e-Science understandings into iterations of the life cycle model allowing scientists to understand the objectives of applied e-Science life cycles.



The intelligence of the proposed e-Science life cycle model lies in its capability as customizable relationship model for scientific dataspaces, as it covers the creation, representation and searching of semantically rich relationships among participants of a dataspace. It enables researchers to find not only relevant primary data in connection with its derived data, but also lot of semantics about what was initially done with the data, such as which data preprocessing methods have been applied, which data mining and analysis models have been used, which result visualizations are available etc. Furthermore, it points to relevant background data, such as descriptions of applied services, models, research domains, etc.

Conclusions

The information recorded by the e-Science life cycle ontology provides rich semantics about the relationship among primary and derived data in e-Science applications. Additionally scientists will retrieve information about the goals specified in e-Science experiments, which domain it corresponds, and whom to contact in case of interest for engaging collaborations, in short, users will understand for what reason a specific e-Science life cycle was applied, which we summarize by the meaning of e-Science Understanding.

Literature cited: [1] Franklin, M., Halevy, A., and Maier, D. 2005. From databases to dataspaces. ACM SIGMOD. [2] Halevy, A., Franklin, M., and Maier, D.. 2006. Principles of dataspace systems. Proceedings of the ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS).
For further information: Please contact elsayed@par.univie.ac.at. More information on this and related projects can be obtained at www.par.univie.ac.at. This poster is available online at www.gridminer.org/e-sciencelifecycle. Funding for this project is provided by the Austrian Federal Ministry of Science and Research.