WHITE PAPER

New Database Engineering and Archive Construction Technology to Accelerate Bio-Imaging, Biomedical Engineering, and Covid-19 Research

LTS (Linguistic Technology Systems) is founded by Amy Neustein, Ph.D., Series Editor of **Speech Technology and Text Mining in Medicine and Health Care** (de Gruyter); Editor of **Advances in Ubiquitous Computing: Cyber-Physical Systems, Smart Cities, and Ecological Monitoring** (Elsevier, 2020); co-author (with Nathaniel Christen) of **Cross-Disciplinary Data Integration and Conceptual Space Models for Covid-19** (Elsevier, 2021); and co-editor of **Medical Image Processing and Machine Learning** (Institution of Engineering and Technology, forthcoming).

Team

Principal Investigator: Dr. James A. Rodger, Professor of Management Information Systems and Decision Sciences at Indiana University of Pennsylvania

Administrative Officer: Dr. Amy Neustein, founder of Linguistic Technology Systems (LTS)

Contributors

- Nathaniel Christen, Lead Software Architect, LTS
- Professor Amita Nandal, Department of Department of Computer and Communication Engineering at Manipal University, Jaipur
- Professor Arvind Dhaka, Department of Department of Computer and Communication Engineering at Manipal University, Jaipur; recently visiting scholar at University of Varna, Bulgaria
- Professor Todor Ganchev, Vice Rector of Research at University of Varna, Bulgaria

---

The "MOSAIC Data-Set Explorer" (MdsX) and "MOSAIC Structured Reporting" (MOSAIC-SR) are tools to help authors develop interactive presentations supplementing academic documents (MOSAIC is an acronym for "Multi-Paradigm Ontologies for Scientific and Technical Publications"). With MdsX, interactive presentations take the form of software applications that provide access to data sets, analytic techniques, or other digitally representable artifacts to document or encapsulate research work. With MOSAIC-SR, authors can implement or reuse code libraries that report on research/experiment methods, workflows, and protocols. Conceptually, MOSAIC-SR is functionally similar to the various domain-speciic recommendations collectively gathered into the "Minimum Information for Biological and Biomedical Investigations" (MIBBI) specifications, and indeed one use-case for MOSAIC-SR is that of implementing object models instantiating MIBBI policies. In some contexts, MOSAIC-SR and MIBBI overlap, because elements of scientific workflows are sometimes algorithms implemented within a code package concretizing authors' research.

MOSAIC-SR can express both computational workflows that are fully encapsulated by published code as well as real-world protocols concerning laboratory equipment and physical materials or samples under investigation. In the latter guise, MOSAIC-SR code can employ or instantiate standardized terminologies and data structures for describing experiments — such as MIBBI policies or BIOCODER functions. In this case, the role of MOSAIC-SR code is to serve as a serialization/deserialization endpoint for sharing research metadata. Conversely, when workflows are fully implemented within software developed as part of a body of research, MOSAIC-SR can provide a functional interface allowing this code to be embedded in scientific software. For these cases, MOSAIC-SR

provides a framework for modeling how a software component specific to a given research project exposes its functionality to host and/or networked peer applications. There are also scenarios where both scenarios are relevant — the MOSAIC-SR code would simultaneously document real-world experimental protocols and construct a digital interface as part of a workflow which is part digital and part "real-world."

This paper will focus on one specific application of MOSAIC-SR in the context of image analysis and bioimaging — specifically, a "Data Structure Protocol for Image-Analysis Networking" (D-SPIN), which both extends and adds a narrower focus to MOSAIC-SR. Software using the D-SPIN protocol provides a description of image processing capabilities which have been utilized and/or are functionally exposed by code and data associated with a research project. This includes "structured reporting" of research objectives as well as a concrete interface for invoking analyses associated with the research (either new algorithms or techniques used to obtain reported findings). D-SPIN, in turn, is based on CaPTk (the Cancer Imaging and Phenomics Toolkit) and Pandore (an image-processing environment which includes both data models and interactive software). The Pandore project encompasses an ontology of "Image Processing Objectives" that provides a structural basis for D-SPIN. For information about how different objectives are merged into workflows, D-SPIN adopts protocols from CaPTk, particularly with respect to implementating image-analysis capabilities as extensions to a core application, and CaPTk's implementation of the Common Workflow Language (CWL). In effect, D-SPIN formalizes the data models and prototypes adopted by Pandore and CaPTk so as to concretize MOSAIC-SR for the specific domain of image processing and Computer Vision. The following sections will therefore outline D-SPIN features in the context of MOSAIC-SR design principles and objectives.

Meta-Procedural Modeling in D-SPIN and MOSAIC-SR

Most approaches to modeling research workflows involve some concept of "meta-objects",[1], "tools" (in the terminology of CWL), and "transitions" (in the language of Petri Net theory). In , the analogous concept is that of *meta-procedures*, which are analogous to ordinary computational procedures but add extra sources of information concerning input and output parameters. In general, rather than simply passing an imput value into an executable routine, metaprocedures define steps which can be taken to acquire the proper values when needed. Aside from ordinary runtime values, the most important input sources are methods defined on GUI components; command-line parameters; file contents; and not-yet-evaluated expressions (perhaps encapsulated in scripts or function pointers). A meta-procedure formulation abstracts the acquisition of inputs (or "channels") from the concrete procedure or procedures which are eventually executed. Therefore, a MOSAIC meta-procedure definition has two separate parts: a preamble where input sources are described, and an executive sequence where concrete procedures are indicated. A *meta-evaluator* then operates in accord with these definitions, concretizing the input values and launching the actual procedure(s). For D-SPIN, meta-procedures can be defined using a framework based on BioCoder,[2] but adopted to the imaging and Computer Vision context.

Image analysis methods are often described in academic literature in terms of methematical formulae and/or characterizations of computational environments (such as Graphical Processing Units); it requires additional construction to translate these overviews into actual computer code. Once Computer Vision innovations are in fact concretely implemented, there is then an additional stage of development requisite for users to actually enact the computations described in the research. Although it is theoretically possible to demonstrate novel methods within fully self-contained autonomous applications, it is more convenient for users if research code is integrated with existing imaging software. The D-SPIN interface can then help connect new code to existing applications, allowing users to access new code's functionaloty through GUI actions, command-line

[1] See the VISSION system: https://pdfs.semanticscholar.org/1ad7/c459dc4f89f87719af1d7a6f30e6f58dff17.pdf.
[2] See https://jbioleng.biomedcentral.com/articles/10.1186/1754-1611-4-13

invocations, or inter-application messaging protocols.

In addition to practically enabling application embedding, D-SPIN models represent research methods and theories, contributing to transparency and reusabiltity according to the MIBBI and FAIR (Findable, Accessible, Interoperable, Reusable) standards.[3] This can be achieved, in part, by implementing data structures conforming to Image Processing Objectives. However, D-SPIN embeds this logic in an Object-Oriented context which allows imaging-specific workflow notations to be paired with specifications outside of imaging processing in the narrowest sense. This allows D-SPIN to be available for hybrid computational objectives representations which are only partially covered by the imaging domain — analogous to the MIAPE-GI (Gel Electrophoresis Informatics) component of MIAPE (Minimum Information About a Proteomics Experiment). The following section will discuss several domains where D-SPIN has been explicitly integrated with code libraries codifying MIBBI-style research protocols.

D-SPIN in Contexts Supplemental to Image Processing.

Image Flow Cytometry

One important use-case for biomedical image processing is to analyze cellular microscopy in conjuncton with cytmetric experiments which investigate cells and cellular-scale entities (such as proteins) indirectly. Conceptually, image analysis and flow cytometry (FCM) analysis are mathematically similar, and some commercial cytometry software has been extended with image-processing capabilities. The overlap between cytometric and image analysis has also inspired attempts to merge cytometry standards, such as MIFlowCyt (the Minimum Information about a Flow Cytometry Experiment policy within MIBBI), with bioimaging standards such as DICOM (Digital Imaging and Communications in Medicine). One such proposal is due to Robert Leif, who argues that "The large overlap between imaging and flow cytometry provides strong evidence that both modalities should be covered by the same standard" and has formalized an XML language (CytometryML) to serve as that overarching bridge. The D-SPIN project builds off this work by introducing its own FCM/DICOM hybrid, although in an object-oriented rather than XML-based context (discussed further in the next section). As a reference implementation for this D-SPIN extension, the project also provides a pure-C++ cytometry library based on the openCyto and libraries, but eliminating external dependencies such as R and Java. The FCM/DICOM bridge is implemented in this context via a D-SPIN supplement to DICOM based on "Semantic DICOM," which is an effort to standardize query processing within PACS (Picture Archiving and Communications Service) workstations and to more effectively integrate DICOM with clinical data.

A Semantic DICOM Object Model

[3] See https://www.researchgate.net/publication/331775411_FAIRness_in_Biomedical_Data_Discovery.