The datamodel: key to developing a SAS data transfer framework

Hester, J. R.[1]

[1] *Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia*

SUMMARY

"Data" in a broad sense includes both the core measurements and the descriptive information surrounding those measurements. In order to transfer data effectively, both parties must agree on (I) the electronic representation of the information (commonly referred to as the "data format") and (II) the precise meanings and attributes assigned to the data items encapsulated in the format. While the former requires little more than choosing from any number of off-the-shelf format specifications, the latter requires broad agreement on unambiguous definitions. At the interface between the data format of (I) and the data definitions of (II) lies the data model: this is simultaneously the abstract data structure produced by the data format parser, and the abstract data description that the authors of definitions must link back to the scientific domain. A complex data model will deter members of the scientific community from participation in definition development. Conversely, an overly simple data model will not provide sufficient expressive power to allow computer-driven data manipulation, and may lack flexibility to handle inevitable future changes. The "sweet spot" lies at the simple end of the complexity spectrum where data models describe little more than collections of key-value pairs or collections of data tables. The currently popular hierarchical data formats encourage complexity for little, if any, gain.

Data transfer protocols benefit from a set of standards for judging dataset quality. The standards provide ready-made data definitions, and journals and referees can require that submissions meet the standards, driving adoption of a data framework that can incorporate information required by the journals at each stage of data processing.

Requirements for a successful SAS data framework therefore become: (i) a set of standards for SAS data; (ii) a datamodel that is just complex enough to efficiently express this data and potential future data; (iii) one or more formats that can be transformed to and from the target datamodel.