**WHITE PAPER**

# A Novel Scientific Data Repository Framework

The Scientific Data Repository Framework (SDRF) encompasses data models, publishing guidelines, and code libraries to help authors and publishers deploy open-access research data sets that are associated with scientific publications. Nowadays there are many general-purpose and domain-specific portals hosting scientific data; there are also several available formats for describing and encoding scientific data, such as Research Objects, schema.org/Dataset, Digital Curation Center, SciData, BioCoder, and MIBBI (Minimum Information for Biological and Biomedical Investigations). However, SDRF is unique in merging different data-set formats into a unified, overarching standard which can be adapted to different publishing environments.

In order to conform to current specifications — such as FAIRsharing (Findable, Accessible, Interoperable, Reusable) or the Bill and Melinda Gates Foundation guidelines for authors — proper protocols must be implemented at several stages of the publishing process. In particular, (1) publications themselves should provide clear descriptions of accompanying data sets and where that data is hosted; (2) publication repositories (such as Springer Nature) should make data-set links and metadata clearly visible on web pages where documents are read or previewed; (3) data sets themselves need to include metadata and supporting files which help researchers properly access, visualize, and reuse the data; and (4) data sets need to be connected with software which has the correct features to load and display the relevant raw data files. SDRF will include technology applicable to each of these four facets of the publishing and data-sharing pipeline.

It is important to emphasize that data sets are only truly valuable if they are machine-readable and seamlessly integrated into domain-specific software ecosystems. Scientists who carefully examine and reuse published data sets are usually researchers doing technical work in a field closely related to the original authors'; in many cases there are specialized software applications, computational methods, algorithms, and research protocols which are endemic to the relevant subject areas. When sharing research data, accordingly, publishers and authors should make it as easy as possible for scientists to examine the data within the digital ecosystem they utilize for their own research. This often means that document viewers — e.g., PDF viewers and/or HTML pages on publisher portals — should ideally be interconnected with scientific applications, so that scientists, when reading books/articles, can seamlessly launch domain-specific software and visualize/examine associated data sets. Unfortunately, most scientific software does not incorporate code libraries to parse metadata describing open-access data sets. Therefore, publisher and data-hosting repositories can only be truly interoperative with scientific applications by providing plugins or extensions that add data-set-accession capabilities to existing scientific software.

To demonstrate how such plugins can work, as well as other facets of the data-publication process augmented via SDRF, this paper will review two case-studies involving existing or forthcoming articles.

## First Case Study: "Parkinson's Disease Diagnosis: Effect of Autoencoders to Extract Features from Vocal Characteristics" from the International Journal of Speech Technology, forthcoming

This case study demonstrates a scenario where one article reuses multiple pre-existing data sets. The article, by Ashena Gorgan Mohammadi, Pouya Mehralian, Amir Naseri, and Hedieh Sajedi, examines Parkinson's Disease symptomology from multiple perspectives, including gait (loss of motor function), speech impairment, and bioimaging (MRIs). The authors apply Machine Learning to data sets focused on these three different diagnostic areas, advansing research to refine Parkinson's predictors and diagnoses. The overall information can be summarized as follows:

1. Gait data was primarily drawn from a PhysioNet data set (https://physionet.org/content/gaitpdb/1.0.0/) obtained via sensors attached to subject's feet as they walked (the study includes both Parkinson's patients and healthy controls). This sensor data is provided as a collection of text files, each file corresponding to one patient (or control subject), with each line in a file representing a single time snapshot. The lines are divided into space-separated columns each representing force exerted on a single sensor (plus two columns for total force on the left-foot and right-foot sensors respectively). This data set also includes demographic and clinical information for each patient, in a spreadsheet format.

   An additional source of gait data used by the authos is a more recent (2019) study whose data is shared only on request (see https://www.nature.com/articles/s41598-019-53656-7#MOESM1).
2. The primary source of speech data is a data set hosted by the University of California Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Parkinsons). The central information is a CSV file, where each line represents a single voice recording from one of 23 Parkinson's patients or 8 healthy controls (each subject made multiple recordings). The individual lines present a quantitative model of subjects' speech via a collection of acoustic features/attributes. A similar "telemonitoring" data set from the same archive is used to study how analysis of Parkinson's-related speech data may carry over to samples obtained via devices such as smartphones, for "people who may have difficulties maintaining a precise clinical examination routine."
3. Radiological data for the Parkinson's analysis is not immediately available for reuse, but requires special (non-commercial) authorization from the Parkinson's Progressive Markers Initiative (https://www.ppmi-info.org/) or corresponding authors of papers introducing the relevant data (https://www.frontiersin.org/articles/10.3389/fnins.2019.00874/full#h6). Although this data is derived from bioimaging — so that the underlying raw data files are radiological — the current authors utilize the data in a more structured form, building off of image-feature extraction already performed during the prior cited research. However, the republished unified data set will include code allowing researcher to reproduce the original analytic workflow if desired; for instance, we will enable the implementations published via https://www.frontiersin.org/articles/10.3389/fnins.2019.00874/full#h6 to be embedded as a CaPTk module (see in particular https://cbica.github.io/CaPTk/tr_integration.html#tr_cppIntegration).
4. The authors also provide a Python code repository, hosted on GitHub, which contains code they used to analyze these various data sources.

In their bibliography, the authors cite the data sets they used either directly or by referencing publications where the relevant data sets are described. This properly credits authorship to the researchers who curated the data sets, and it gives readers a means to locate the raw data. However, accessing and working with the raw data is inconvenient from a reader's point of view without most or all of the data sets being repackaged into a single archive that could be hosted and downloaded as a unit. Obtaining raw data from the resources identified in the bibliography requires several steps — for instance, the PhysioNet sensor data can be downloaded as a zipped folder, while the demographic data attached to it has to be downloaded separately. Some of the information obtained from MRI and speech analysis (reported in papers cited as data sources for the primary article) is provided as supplemental materials within the secondary papers, which requires readers to browse through the articles so as to find downloadable links. In short, piecing all the source data together forces readers to manually inspect multiple web resources and to manually interconnect files once they are downloaded.

Another complicating factor is that certain information present in the data sets is implicit within how the data sets are organized, requiring extra effort to extract this information in a machine-readable manner. For instance, the PhysioNet sensor data uses a file-naming convention which encodes several pieces of information in the file names, such as whether the file presents a Parkinson's patient or a control subject. By examining file names, it would be possible to construct a table with additional information providing context for the file contents. However, such information is not directly included

within the PhysioNet data set; it needs to be extracted by computer code.

This collection of data sets serves as an example of how technologies such as SDRF can make scientific data more "FAIR" (Findable, Accessible, Interoperable, Reusable). When the forthcoming article is published in the International Journal of Speech Technology, the disparate open-access data from the article's secondary sources will be provided as a single SDRF archive. This archive will provide machine-readable access to information spread across multiple sources, translated into a common file-format. In general, SDRF encourages and implements features to help data sets conform to FAIR and related standards, such as (1) bundling multiple data sets into a single archive; (2) migrating data to general-purpose representations wherever possible — formats such as XML, HDF5, ARFF, or DICOM; (3) providing meta-data in several formats, such as Research Objects, schema.org/Dataset, Digital Curation Center, SciData, BioCoder, and MIBBI (Minimum Information for Biological and Biomedical Investigations); (4) identifying one or more "preferred applications" for examining/reusing the published data; (5) explicitly representing information encoded via file-names; (6) bundling raw data, meta-data, and (where possible) machine-readable article text into a single resource, which SDRF calls a "Supplemental Archive;" and (7) annnotating the data sets to support microcitations granularly linking the publication to its associated Supplemental Archive.

An important question for any SDRF archive is how researchers will productively access the data once a Supplemental Archive has been downloaded. Unlike the Flow Cytometry use-case discussed next, the Parkinson's archive spans several scientific disciplines, and there is no obvious application which could be preferred by default for examining the data files. As a fallback option, SDRF is designed to present data sets via Qt Creator, a C++ Integrated Development Environment associated with the Qt application-development framework. SDRF includes code libraries to represent research meta-data as C++ objects, and these libraries can be opened as Qt projects. These may be supplemented with separate libraries extracting and managing information specific to individual data sets. In particular, the Supplemental Archive for the forthcoming Parkinson's article will provide C++ classes encapsulating spreadsheet-like data (whether originally in **.xls**, , or space-delimited formats) republished in the unified data set.

An additional concern for SDRF archives is how to properly annotate publications and data sets side-by-side. In the Parkinson's article, individual C++ classes encapsulating tabular data serve as convenient microcitation targets: annotations within the relevant C++ code represent anchors through which the data set may be referenced (on a more precise scale than merely referencing the Supplemental Archive as a whole). In some cases, individual class attributes can also be linked to lines in the authors' Python source code. On the text side, certain individual paragraphs within the Parkinson's article specifically discuss individual data sets that the authors analyzed, the same data which the SDRF archive encodes via C++ classes. Therefore, those paragraphs can be cleanly linked to the corresponding C++ code annotations. This illustrates SDRF's recommended annotation/microcitation system, where segments in publication texts (identified for instance via LaTeX **phantomsection** commands or JATS **statement** tags) are linked to annotations or comments in code and/or raw data files in the Supplemental Archive.

## Second Case Study: "Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia" from nature.com, 2020

This article presents a use-case with some noteworthy contrasts to the Parkinson's publication described in the previous section. The Covid-19 paper (https://www.nature.com/articles/s41467-020-17292-4) was published along with two data sets comprising Flow Cytometry Standard (FCS) files hosted via the Flow Repository (http://flowrepository.org/id/FR-FCM-Z2N4 and http://flowrepository.org/id/FR-FCM-Z2N5). Links to the data sets (via **flowrepository.org** pages) are explictly provided in the publication's "Data Availability" section. However, researchers still need to perform several steps to manually download the full set of relevant FCS and meta-data files.

One feature of this second use-case is that the technical information in the data sets belong to a

single scientific area (Flow Cytometry) and are encoded in a single format (FCS). As such, it is straightforward to identify the kind of software which researchers need to use to visualize the raw data — any application which can parse FCS files. There are a variety of commercial as well as open-source Flow Cytometry applications which can be used to access FCS data. Once readers have downloaded the Flow Repository archives, then, they may individually load the **.fcs** files to study data which, in the original article, is summarized via figure illustrations.

This workflow nonetheless requires researchers to perform several manual steps before being able to use the published data. The core problem is that existing Flow Cytometry software does not intrinsically have capabilities to read PDF files, locate FCS data sets, and interoperate with hosting platforms such as Flow Repository. Employing this use-case as an example with which to demonstrate proper alignment between document viewers, publication/date repositories, and scientific software, Linguistic Technology Systems is developing a new Flow Cytometry application which *can* interoperate with PDF viewers and SDRF archives. This application is designed so that, when authors are reading a PDF file associated with an FCS data set, the PDF viewer can automatically launch and signal to the FCM application when a reader wishes to download and visualize FCS files. In short, the FCM application — having received data from a PDF viewer which implements an SDRF inter-application protocol — will automatically preform download and extraction steps that scientists otherwise would have to perform manually.

Note also that, although most of the relevant data for this Covid-19 article is in FCS form, there is also supplemental clinical information provided in other formats. For cases such as these, the LTS FCM software includes code libraries allowing researchers to parse non-FCS data as well in standard formats such as XML, HDF5, or ARFF.

This use-case illustrates a general principle, that research data is most convenient for scientists when it is deployed within an infrastructure where portals, document viewers, and scientific applications may seamlessly interoperate. Wherever possible, when they are reading published books or articles, researcher should be able to automatically launch the proper scientific application, download data sets, and examine raw data files in the preferred software with only one or two clicks — instead of manually finding, downloading, merging/extracting, and then opening data files, these steps should be performed automatically as much as is feasible.

## Conclusion

The two use-cases considered here are similar in that each concern articles which are linked to multiple data sets, and for maximum convenience it is optimal for researshers to be able to access this data without performing manual download andd merging/extracting actions. There are also some differences: in particular, the Parkinson's data spans multiple disciplines, whereas the Covid-19 data is more rigorously grounded in Flow Cytometry. As such, the operational requirements for the Covid-19 data, from a reader's point of view, are more clearly delineated: effective integration between the publication and its concomitant data sets is defined by launching Flow Cytometry software while a researcher is reading the publication, allowing the researcher to view the data via software similar to that used to generate/analyze the data while the reported research was being conducted. In the case of the Parkinson's data, there is no single application which would seamlessly display the specrum of information considered in that article; as mentioned above, SDRF defaults to using QT Creator as a fallback for loading Supplemental Archives where no other software is available.

Whether using QT Creator or a domain-specific application, it is preferable that each SDRF archive be associated with one or more applications that researchers can use to view data (and extract information) from the archive. Moreover, these applications would ideally be linked to document viewers and also to publisher's portals, so that readers can automatically launch preferred applications and view Supplemental Archives while reading concordant publications. In order to achieve this, scientific applications need to be augmented with plugins to parse SDRF data. In addition, we are developing an inter-application messaging protocol so that disparate applications with SDRF plugins

may interoperate. In particular, PDF viewers may interoperate with scientific applications so that publications' data sets may be automatically downloaded and visualized via the preferred software.

Our prototype example for an application utilizing such plugins, as mentioned above, will be software for Flow Cytometry. We are also working on a Qt Creator plugin so that Qt Creator (as the "default" SDRF software) can participate in SDRF networks using the same protocol. We will then expand the scope of this protocol via plugins for software concerning implementation such as image-analysis, molecular visualization, radiology, 3D graphics, and so forth.