Data Packaging Formats for research data

This document gives a rough list of projects and tools which help with structuring research data and related outputs into portable, preferably self-describing “packages”.

Note (2017-11-03): this document is currently being re-formulated into a table for greater consistency etc., and will probably be transferred to github or wikipedia at some point. Where present, the information in the spreadsheet will be more accurate/up-to-date. For now, please make suggestions or edits here or in the spreadsheet version: https://docs.google.com/spreadsheets/d/1Tg-oYGPdBDs5LORt0olD5t4X1R_YliUrwr6bNmImjTk/edit#gid=0

Definition of “Data Packaging Formats”

TODO - characteristics, properties of formats/projects included in the list

Some preliminary thoughts:

- "self-describing"/”self-documenting”: metadata bundled with content
- offline first: like a git repository, the package is stand-alone and does not need any out-of-band information, e.g. no need to be able to connect to a central server
- portability: stand-alone packages work equally well regardless of where they are stored/managed and can be moved between locations without loss of information, i.e. stored on local disk, downloadable over HTTP, serialised/modelled according the the web architecture e.g. LDP, sent to object storage, etc.
- minimally prescriptive: e.g. in terms of the arrangement of actual content (with the possible exception of adding an outer wrapper folder as is done with Bagit)
- can interoperate with other tools that operate with a folder as the unit if management, e.g. git, bagit, archive formats like zip, tar, etc.
- The packaging format should not be data-format sensitive
- The packaging format should not be research domain specific
- The packaging format should not be technology or platform specific
- The data package should contain as much contextual information as possible
- Metadata should be easily human and machine readable
- The package format should contain self-checking and verification features

Use-cases for Data Packaging Formats

TODO

List of Data Packaging Formats/Projects (currently active)

Data Packages/Frictionless data

- http://frictionlessdata.io/data-packages/
- Has a manifest, metadata is in JSON, uses its own vocab
- Extensions allowed
- See discussion re json-ld https://github.com/frictionlessdata/specs/issues/110
- See discussion re metadata namespaces https://github.com/frictionlessdata/specs/issues/403
- C.f. w3c CSV on the Web recommendation
- Discussion of spatial data packages: https://research.okfn.org/spatial-data-package-investigation/
- Discussion of adoption by OpenML (open machine learning) in addition or as a replacement for their existing ARFF package convention: https://docs.google.com/document/d/1c_RhDiXTK5bEsY5gGRuQwaF6fKilt4jKq2c_BRqyEDc/edit#heading=h.e7kj1g6iptel
- On FAIR Data: http://blog.okfn.org/2018/08/14/frictionless-data-and-fair-research-principles/

Research Objects

- http://www.researchobject.org/
- Abstract model with ORE manifest serialized as any RDF, but serialization as bundle.zip and bagit have manifest in JSON-LD
- Wikipedia page: https://en.wikipedia.org/wiki/Research_Object (not exclusively about researchobject.org but clearly largely based on it)
- Bagit Research Object: This GitHub repository explains by example a profile for a BagIt bag to also be a Research Object. - https://w3id.org/ro/bagit - used by http://bd2k.ini.usc.edu/tools/bdbag/
- See also the following projects which use ResearchObjects:
    - BioCompute Object (BCO) https://osf.io/h59uh/
    - BDBags http://bd2k.ini.usc.edu/tools/bdbag/

DataCrate:

- https://github.com/UTS-eResearch/datacrate/tree/master/spec/0.1
- A method of packaging research data based on BagIt with additional metadata in a **human-and-machine-readable** format based on schema.org linked-data
- Uses Bagit + HTML/rdfa manifest with Schema.org metadata, datacite.xml where available
- File-level metadata support
- Tool: https://codeine.research.uts.edu.au/eresearch/calcyte
- http://ptsefton.com/2017/10/19/datacrate.htm
- http://cameronneylon.net/blog/packaging-data-the-core-problem-in-general-data-sharing/

- This specification is a practical guide for software authors to create tools for generating and consuming research data packages. The format is **not optimized for hand-authoring**, it is optimized for maximum convenience for data consumers with the use of HTML for users to read, for search engines with RDFa metadata embedded in the HTML and JSON-LD for programmers.
- In an academic context, it is important to be able to associate data with funded research projects and publications, and desirable to be able to describe the *contents* of a data package, not just the data in aggregate, so DataCrate allows for file-level descriptions; to describe licensing, copyright ownership and authorship and other metadata at a granular level.
- Example DataCrate: `https://doi.org/10.5281/zenodo.844394`

RDA Repository Interoperability Package

- http://dx.doi.org/10.15497/RDA00025
- The definition of an exchange format and the description of functional requirements needed in order to exchange data in the defined format.
- Bagit

Darwin Core Archive

- http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102623
- http://www.gbif.org/resource/80636
- https://github.com/gbif/ipt/wiki/howToPublish
- http://tools.gbif.org/dwca-assistant/
- http://wiki.ggbn.org/ggbn/DwC GGBN Data Standard and Darwin Core-Archive
- Wikipedia page: https://en.wikipedia.org/wiki/Darwin_Core_Archive

DataOne Data Package

- ORE + bagit based
- Includes ProvONE provenance metadata for relationships among components and packages
- https://releases.dataone.org/online/api-documentation-v2.0.1/design/DataPackage.html
- https://www.rdocumentation.org/packages/datapack/versions/1.0.1/topics/datapack
- See also, discussion of Dryad data exchange packages based on ORE + bagit: http://wiki.datadryad.org/images/e/e2/Ksclarke-c4l11.pdf
- Extension/implementation from hydrology domain: http://digitalcommons.usu.edu/cee_facpub/1182/
- Also seems to be used by sead-data http://sead-data.net/about/standards
- Downloadable examples
    - Example with provenance: https://arcticdata.io/catalog/#view/doi:10.18739/A2556Q
    - https://knb.ecoinformatics.org/#view/doi:10.5063/F19021QT
- Wikipedia page: https://en.wikipedia.org/wiki/DataONE (parent project rather than packaging spec)

Dataset Publishing Language (DSPL)

- https://developers.google.com/public-data/overview
- Google initiative
- CSV Files only
- Wikipedia page: https://en.wikipedia.org/wiki/Google_Public_Data_Explorer (parent project of DSPL)

Web Packaging Format

- https://github.com/WICG/webpackage
- Continuation of https://www.w3.org/TR/web-packaging/
- IETF & W3C (WICG)
- Single file (application/package) with option compression
- Allows for but doesn't prescribe a metadata file (config.xml)
- Overview/Explanation: https://github.com/WICG/webpackage/blob/master/explainer.md
    - See also w3c App Manifest work: https://www.w3.org/TR/appmanifest/

CSV on the Web

- https://www.w3.org/2013/csvw/wiki/Main_Page
- https://www.w3.org/TR/2015/REC-tabular-data-model-20151217/
- Tabular data is routinely transferred on the web in a variety of formats, including variants on CSV, tab-delimited files, fixed field formats, spreadsheets, HTML tables, and SQL dumps. This document outlines a data model, or infoset, for tabular data and metadata about that tabular data that can be used as a basis for validation, display, or creating other formats. It also contains some non-normative guidance for publishing tabular data as CSV and how that maps into the tabular data model.
- The mission of the CSV on the Web Working Group, part of the Data Activity, is to provide technologies whereby data dependent applications on the Web can provide higher interoperability when working with datasets using the CSV (Comma-Separated Values) or similar formats. As well as single CSV files, the group will define mechanisms for interpreting a set of CSVs as relational data. This will include the definition of a vocabulary for describing tables expressed as CSV and locatable on the web, and the relationships between them.
- C.f. Frictionless Data Tabular Data Format

Data Conservancy Packaging Specification

- ORE + Bagit
- http://dataconservancy.github.io/dc-packaging-spec/
- https://wiki.library.jhu.edu/display/DCSDOCPKG/4.+DC+Packaging+Specification
- http://dataconservancy.org/wp-content/uploads/2016/10/DataConservancyComponents.pdf
- http://www.dcc.ac.uk/sites/default/files/documents/IDCC15/196.pdf
- https://github.com/DataConservancy/dcs-packaging-osf
- https://www.youtube.com/watch?v=uLRNVbsLAok
- http://dataconservancy.org/data-conservancy-and-data-rescue-boulder-pilot/

Research Compendium (Packaging research data as R/rstats packages)

- https://rawgit.com/benmarwick/MacquarieRusers-2017/master/Marwick-MacquarieRusers-2017.html#1
- Examples:
  - https://github.com/cylerc/AP_SC
  - https://github.com/benmarwick/1989-excavation-report-Madjebebe
  - https://github.com/benmarwick/ktc11
- See also
  - https://sjfox.github.io/post/2017-05-04-rtzikvrisk_primer/?utm_content=bufferdae35&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer
  - https://github.com/ropensci/rrrpkg
  - http://o2r.info/erc-spec/spec/#r-workspaces
  - https://github.com/jhollist/manuscriptPackage
  - https://github.com/cboettig/template
  - https://github.com/rubenarslan/codebook
    - Automatic Codebooks from Survey Metadata Encoded in Attributes
  - dataspice (see below)

Executable Research Compendium (ERC)

- http://o2r.info/erc-spec/spec/
- http://www.dlib.org/dlib/january17/nuest/01nuest.html
- “An Executable Research Compendium (ERC) is a packaging convention for computational research. It provides a well-defined structure for data, code, text, documentation, and user interface controls for a piece of research and is suitable for long-term archival. As such it can also be perceived as a digital object or asset.”
- Base directory with ERC confguration file
- Must include concept of a ‘main’ file, which is a literate programming file intended to represent the executable component of the package (e.g., main.Rmd)
- Must include a executable runtime image and mainfest
- Serialized using BagIt
- https://docs.google.com/presentation/d/1RE6w6yTo0FMAwIF8YitV6hnivRVvN5VDFdK9nrfzwtk/edit

Metatab Packages

- https://github.com/Metatab/metatab-declarations/blob/master/specs/Metatab%20Packages.md
- http://metatab.org/
- The standard proposed here is very simple. In the simplest case, for data in Excel format, it involves adding an additional worksheet, named ‘meta’, which holds information about the name, title and creators of the data, and another worksheet called ‘schema’ with a list of the data columns and their data types and descriptions.

This proposal defines similar package structures for ZIP archives, file systems and web pages

- Note from Paul Walsh, Frictionless Data: “Metatab is really a superset of Tabular Data Package. We have active discussions with Eric at Metatab on moving towards the "meta" tab becoming a core supported serialisation of datapackage.json”

ReproZip

- https://www.reprozip.org/
- ReproZip allows you to pack your research along with all necessary data files, libraries, environment variables and options.Then anybody can reproduce the research on a different machine, without tracking down and installing the dependencies, or even having to run the same operating system!
- No descriptive metadata like datacite? Focused mostly on code.

Science.ai archive format

- https://nightly.science.ai/documentation/archive
- https://lists.w3.org/Archives/Public/public-scholarlyhtml/2017Oct/thread.html

COMBINE archive format

- http://co.mbine.org/documents/archive
- https://www.ncbi.nlm.nih.gov/pubmed/25494900
- A COMBINE archive is a single file containing the various documents (and in the future, references to documents), necessary for the description of a model and all associated data and procedures. This includes for instance, but not limited to, simulation experiment descriptions, all models needed to run the simulations and associated data files. The archive is encoded using the Open Modeling EXchange format (OMEX).

Codemeta https://codemeta.github.io/

- Minimal metadata schemas for science software and code, in JSON-LD
- https://doi.org/10.5063/schema/codemeta-2.0
- https://github.com/codemeta/codemeta

IPFS-Pack

- https://github.com/ipfs/ipfs-pack
- Defines a discrete dataset with ipfs
- Compatible with bagit (similar manifest)
- Possibly compatible with data-packages for metadata/semantics
- Content-addressed, global peer-to-peer distribution
- In use to backup and distribute data.gov and other public datasets

Portable Encapsulated Projects (PEP)

- https://pepkit.github.io

- “Portable Encapsulated Projects (PEP for short) is a standardized way to organize metadata. By project, we mean a collection of data, and PEP provides a standard way to describe that dataset. It was designed for use in bioinformatics projects that involve lots of data for a set of samples (which could be individual experiments, organisms, cell lines, what have you). However, the concepts are fundamentally generic and could be applied to any type of project that can represent metadata in tabular form, with rows corresponding to units of interest (i.e. samples) and columns corresponding to attributes of those units.”

Dataspice

- https://github.com/ropenscilabs/dataspice
- R library rstats
- The goal of dataspice is to make it easier for researchers to create basic, lightweight and concise metadata files for their datasets.
- json-ld with schema.org
- Built around a configurable folder convention, e.g. /data /docs directory
- Uses csv templates to generate json-ld

Pysch-DS

- A specification for psychological datasets
- Psych-DS incorporates a few types of existing recommendations for organizing our work (well-formatted spreadsheets, data dictionaries, and sensible folder structure) into a technical specification - a series of requirements for file structure and format that constitute a standard machine-readable template.

WholeTale Format

- Spec: https://docs.google.com/document/d/11mHpqByI1odaRydoz3mSdXZQ6eBwg1JVzPfcCKejyO4/edit#
- Part of the wider wholetale project https://wholetale.readthedocs.io/en/stable/README.html

BIDS

- http://bids.neuroimaging.io/
- The BIDS format is essentially a way to structure your data / metadata within a hierarchy of folders. This makes it easy to browse from a computer, as well as to automatically parse a BIDS folder with a program. The BIDS structure makes minimal assumptions about the tools needed to interact with the data that's inside.
- http://dx.doi.org/10.1038/sdata.2016.44
    - The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments
- Spec: http://bids.neuroimaging.io/bids_spec.pdf

Other possible candidates

Other specifications which require further examination. Share some elements of research data packages

Cross-Linguistic Data Formats

- https://cldf.clld.org/
- To allow exchange of cross-linguistic data and decouple development of tools and methods from that of databases, standardized data formats are necessary
- Based on W3C's CSV on the Web; json-ld

Oxford Common File Layout (OCFL) Objects

- https://ocfl.io/draft/spec/
- "A conceptual gathering of all files (data and metadata), the directories they are organized in, and their changes over time which together form the digital representation of an entity that need to be managed, in preservation terms, as a single coherent whole (i.e., content);"
- "A file and directory layout and administrative information on a storage medium that provides a defined structure for the storage of this content, and through which these files and their changes may be understood (i.e., structure)."
- Slightly different, but related use-case.
  - Is silent about what metadata is included in an OCFL Object (only that it should include all relevant metadata)
  - You could have an OCFL storage root made up of OCFL Objects which are in turn examples of one or more other packaging formats

Network Common Data Form (NetCDF)

- "self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data"
- https://en.wikipedia.org/wiki/NetCDF

Hierarchical Data Format (HDF5)

- https://en.wikipedia.org/wiki/Hierarchical_Data_Format

The Allotrope Framework

- HDF5 based
- The Allotrope Data Format (ADF) is an innovative federation of standards that features the ability to store datasets of nearly unlimited size and complexity in a single file, organized as a single or multiple n-dimensional arrays to record the measurements of the experiments, including time series and hyper-dimensional data. The ADF is also capable of storing the metadata describing the context of each test and measurement event as well as all the instrument settings. The ADF is portable, allowing easy file transfer and use across operating system and vendor platforms.

- Metadata are stored in the Data Description using an RDF Data Model for process, material, instrument and result details, as well as the metadata describing the Data Cube and Data Package layers, all based on semantic web and linked data concepts (and appropriate W3C specifications).

DCAT Distribution Package

- Use Case and Requirement of the W3C Dataset Exchange Working Group (DXWG)
- “In practice, distributions are sometimes made available in a packaged or compressed format. For example, a group of files may be packaged in a ZIP file, or a single large file may be compressed. The current specification of DCAT allows the package format to be expressed in dct:format or dcat:mediaType but it is currently not possible to specify what types of files are contained in the package.”
- https://www.w3.org/TR/dcat-ucr/#x5-1-dcat-packaged-distributions-id1
- https://www.w3.org/TR/dcat-ucr/#RDIP
- See also use case regarding “Common requirements for scientific data” https://www.w3.org/TR/dcat-ucr/#ID9 https://www.w3.org/TR/dcat-ucr/#ID10 https://www.w3.org/TR/dcat-ucr/#ID11

Freya project (persistent identifiers)

- https://github.com/datacite/freya/issues/2
- Discussion of how to provide direct dataset download (rather than redirection to landing page)
- “only scales with a standard packaging format, and the best candidate in terms of functionality and adoption might be bagit.”
- “recommend to provide the bagit item as "application/zip" content type”
    - From Martin Fenner (Datacite): “We may use a custom content type, such as "application/vnd.bagit+zip".”
- “register as content type "application/zip", expressed as URL that would be https://data.datacite.org/application/zip/10.x/xyz”

Scientific Publication Packages

- Scientific Models: A User-oriented Approach to the Integration of Scientific Data and Digital Libraries
- http://espace.library.uq.edu.au/view/UQ:7942
- http://dx.doi.org/10.2218/ijdc.v1i1.4
- http://doi.org/10.1007/s00799-007-0018-5
- Not clear the extent to which it was adopted or is still active
    - c. 2006/2007 articles
    - Email sent to original authors. No response.

FAIR Data Accessor

- http://www.slideshare.net/markmoby/fair-data-prototype-interoperability-and-fairness-through-a-novel-combination-of-web-technologies
- LDP-based, Linked Data
- Dutch Techcentre for Life Sciences
- Creating a static FAIR Accessor layer over your Zenodo objects
- See also FAIR Assessment and Badges *via DANS* (2017)

SWORD Protocol Packaging Types

- Largely a protocol for deposit/CRUD operations in a repository context
- Has extensive notion of “packages” and “packaging” for deposit purposes
- More SIP rather than AIP
- ? Has defined profiles - SimpleZip, DataBankBagit, DspaceMetsSIP, SCORM Package, MPEG21 DDI Package
- SWORD v 3 is actively being developed (as of November 2017)
- http://swordapp.github.io/SWORDv2-Profile/SWORDProfile.html#packaging
- SWORD Content Package Types version 1.0
    - http://www.swordapp.org/docs/sword-type-1.0.html
      http://www.ukoln.ac.uk/repositories/digirep/index/SWORD_formats
- Wikipedia page: https://en.wikipedia.org/wiki/SWORD_(protocol)
- Examples (Some of these might be broken out as packaging formats in their own right):
    - METSDSpaceSIP/DspaceMetsSIP
        - https://wiki.duraspace.org/display/DSPACE/DSpaceMETSSIPProfile
    - DataBankBagit
        - http://pads.cottagelabs.com/p/databankbagit
        - https://github.com/dataflow/DataStage/wiki/SWORD-overview-for-developers
    - IMS Content Packaging
        - http://www.imsglobal.org/content/packaging/index.html
        - IMS Content Packaging v1.2 Public Draft v2.0 specification describes data structures that can be used to exchange data between systems that wish to import, export, aggregate, and disaggregate packages of content. IMS content packages enable exporting content from one learning content management system or digital repository and importing it into another while retaining information describing the media in the content package and how it is structured, such as a table of contents or which web page to show first. The IMS Content Packaging Specification focuses on the packaging and transport of resources but doesn’t determine the nature of those resources. This is because the specification allows adopters to gather, structure, and aggregate content in an unlimited variety of formats
        - Content Packaging v1.2 is undergoing standardisation by ISO/IEC. We expect the ISO/IEC version to be voted on as a full standard in late November 2009. However, there will be a delay before the ISO/IEC documents are available to buy.
    - IMS Common Cartridge

        - https://www.imsglobal.org/cc/index.html
    - METS
        - http://www.loc.gov/standards/mets/

SCORM Package (for Open Educational Resources)

- Wikipedia page: https://en.wikipedia.org/wiki/Sharable_Content_Object_Reference_Model

Fcrepo Import / Export Format

- https://github.com/fcrepo4-labs/fcrepo-import-export/blob/master/import-export-format.md
- Optionally wrapped in bagit
- Serialisation of Linked Data Platform (LDP) resources to directory structure
- C.f. DSpace AIP Format

DSpace AIP Format

- https://wiki.duraspace.org/display/DSDOC6x/DSpace+AIP+Format
- “AIP is a package describing one archival object in DSpace”
- “Generally speaking, an AIP is an Zip file containing a METS manifest and all related content bitstreams, license files and any other associated files”
- “AIP profile favors completeness and accuracy rather than presenting the semantics of an object in a standard format. It conforms to the quirks of DSpace's internal object model rather than attempting to produce a universally understandable representation of the object. When possible, an AIP tries to use common standards to express objects.”
- “An AIP *can* serve as a DIP (Dissemination Information Package) or SIP (Submission Information Package), especially when transferring custody of objects to another DSpace implementation”
- “AIPs only store the Latest Version of Items. Currently the AIPs that DSpace generates only store the latest version of an Item. Therefore, past versions of Items will always be lost when you perform a restore / replace using AIP tools”
- Interesting because it encompases the idea of packaging the technical & system metadata along with the SIP and DIP metadata
    - “In contrast to SIP or DIP, the AIP should include all available DSpace structural and administrative metadata, and basic provenance information. AIPs also describe some basic system level information (e.g. Groups and People).”
- METS allows for very fine-grained assertions about particular files/bitstreams
- C.f. Fcrepo import/export format

University of Groningen - Dept of Sociology Packages

- http://www.rug.nl/gmw/sociology/research/rdm-sociology-jan-2016-huisstijl.pdf
- Not a technical specification. Departmental recommendation on how to “package”

Distribute papers as R-packages:

- https://sjfox.github.io/post/2017-05-04-rtzikvrisk_primer/?utm_content=bufferdae35&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Jupyter Notebooks

- Json format for encoding rich, interactive documents (ipynb)
- Some conventions around folder structures for location of data etc. :http://jupyter.readthedocs.io/en/latest/projects/jupyter-directories.html
- See discussion here re submission of notebooks to an archive: https://quantecon.org/nb_contrib.html
- Note from Paul Walsh "Various activities in the Jupyter/Pandas ecosystem in the last year suggest that Table Schema / Tabular Data Resource are being used to solve platform-agnostic data package in this space":
    - https://github.com/gnestor/jupyterlab_table
    - https://github.com/pandas-dev/pandas/issues/14386
    - https://pandas.pydata.org/pandas-docs/stable/generated/pandas.io.json.build_table_schema.html
    - https://github.com/nteract/nteract/issues/1526

RIF-CS

- [RIF-CS] is a standard created by the Australian National Data Service based on ISO 2146 for data dissemination which has good general purpose coverage but it's an XML based format that can't be used in a linked-data context directly.
- https://en.wikipedia.org/wiki/RIF-CS
- https://documentation.ands.org.au/display/DOC/Content+Providers+Guide

Cave Archive and Versioning Experiment

- https://suss.caves.org.au/cave/ (referenced here http://ptsefton.com/2017/10/19/datacrate.htm)

Project Open Data (US Gov data and jsonld-based metadata guidelines)

- https://project-open-data.cio.gov/v1.1/schema/
- jsonld-based metadata guidelines, possibly applicable in a packaging scenario

A comprehensive open package format for preservation and distribution of geospatial data and metadata

- https://doi.org/10.1016/j.cageo.2016.09.001

SciDataspace/GeoDatasapce/GeoUnits

- http://web.ci.uchicago.edu/~tanum/SciDataspace/SciDataspace.html
- The GeoDataspace system captures models and data in an integrated way, encapsulates them as a single shareable package, and allows the user to share/publish the package for wider community use or self-preserve it for further analysis.
- https://www.slideshare.net/TanuMalik/geodataspace-simplifying-data-management-tasks-with-globus
- https://daspos.crc.nd.edu/images/workshops/workshop_six/presentations/DASPOS2016a.pdf

- https://www.earthcube.org/group/geodataspace
- Emphasis on reproduction of computational environment, provenance etc
- “Geounit” packages are intended as part of a broader ecosystem

Smart Containers

- Python wrapper around docker which can supplement with json-ld, prov metadata etc
- https://github.com/charlesvardeman/sc_spec
- https://github.com/charlesvardeman/smartcontainers_cfv
- Smart Container (SC) is a modular specification and implementation that leverages linked data principles to enable the preservation, sharing and reuse of both software and data artifacts.
- One community of collaborators at CERN is already exploring Docker to preserve high energy physics experiments. Our approach is to eventually provide a mechanism that captures the additional provenance of computational experiments in a machine readable approach using the W3C standard RDF data model that has been shown to aid in contextualization of scientific experiment.
- The ultimate goal being to provide automated tools that “wrap” the existing Docker command line tool and infrastructure such that it is transparent to the scientist but captures information necessary to populate the metadata behind the scenes.
- Smart Containers relies on this ability to attach metadata to a container or image. Since a label may be an JSON array, Smart Containers attaches a **JSON-LD** serialization of the graph object to the Docker container. All of the previous provenance history of a Docker Object (image/container) is encoded using RDF/OWL based vocabularies
- Last Github commits February/March 2016
- SmartContainers is being developed as part of the Data and Software Preservation for Open Science (http://daspos.org) project. (concluded August 2017 ??)

EPUB

- http://idpf.org/epub

Web Publications for the Open Web Platform: Vision And Technical Challenges

- https://www.w3.org/TR/pwp/

ARFF

- http://weka.wikispaces.com/ARFF
- https://www.cs.waikato.ac.nz/ml/weka/arff.html
- ASCII file format with header information for metadata about data elements
- See discussion re converting to FrictionlessData packages https://github.com/openml/OpenML/issues/482
- http://www.inf.ed.ac.uk/teaching/courses/dme/html/arff.html

W3C Packaging on the Web

- https://www.w3.org/TR/web-packaging/
- Editor’s draft https://w3ctag.github.io/packaging-on-the-web/

- **Superseded by https://github.com/WICG/webpackage (see above)**
- Single file (application/package) with option compression
- Allows for but doesn't prescribe a metadata file (config.xml)
    - See also w3c App Manifest work: https://www.w3.org/TR/appmanifest/

Related Work/Projects

Relevant projects which facilitate some aspect of reproducible research and which may be used in conjunction with the data packaging formats listed above.

Common Workflow Language http://www.commonwl.org/

- The Common Workflow Language (CWL) is a specification for describing analysis workflows and tools in a way that makes them portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments. CWL is designed to meet the needs of data-intensive science, such as Bioinformatics, Medical Imaging, Astronomy, Physics, and Chemistry.

Singularity - http://singularity.lbl.gov

- Singularity enables users to have full control of their environment. Singularity containers can be used to package entire scientific workflows, software and libraries, and even data. This means that you don't have to ask your cluster admin to install anything for you - you can put it in a Singularity container and run

MINIM ontology for scientific data quality assessment:

- http://linkedscience.org/wp-content/uploads/2013/04/paper5.pdf

DAT

- Dat: Distributed Dataset Synchronization and Versioning (May 2017). Max Ogden, Karissa McKelvey, Mathias Buus Madsen.
- https://datproject.org
- Designed ground-up as a distributed, P2P syncing technology for research datasets (like Git meets BitTorrent but for sharing big research datasets).
- Free, open-source. Grant-funded, not-for-profit Code For Science foundation.
        - See also: Paul Frazee: Beaker Browser and crypto secure append-only logs
- See also Dat-in-the-Lab:
    - Wherein Dat lead dev Max Ogden is hanging with Stephen Abrams, one of the most lucid thinker in digital preservation, on awesome 'fieldtrips' to niversity of California campuses to talk with lab techs about syncing their research data workflow with their research data preservation requirements. What good could come from that?
- Dat alongside git: https://github.com/joehand/dat-git-example
    - Kinda like a peer-to-peer solution to git-lfs or gvfs
- Note from Paul Walsh, Frictionless Data: "I've had various discussion with people at DAT about building on DAT with Data Packages and Table Schema, and we (OKI)

are hoping to support this idea with some data-driven research pilots over the next year or so.”

Nanopub - http://nanopub.org/wordpress/?page_id=65

- “Nanopublications are a natural response to the explosion of high-quality contextual information that overwhelms the capacity of conventional research articles in scholarly communication.
- “A nanopublication is the smallest unit of publishable information: an assertion about anything that can be uniquely identified and attributed to its author.
- Individual nanopublications can be cited by others and tracked for their impact on the community.
- “With nanopublications, it is possible to disseminate individual data as independent publications with or without an accompanying research article. Furthermore, because nanopublications can be attributed and cited, they provide incentives for researchers to make their data available in standard formats that drive data accessibility and interoperability.

Stencila - https://stenci.la

- “The office suite for reproducible research”
- https://github.com/stencila/stencila: “Stencila is a platform for creating, collaborating on, and sharing data driven content. Content that is transparent and reproducible, like RMarkdown and Jupyter Notebooks. Content that can be versioned and composed just like we do with open source software using tools like CRAN and NPM. And above all, content that is accessible to non-coders, like Google Docs and Microsoft Office.”

ISA Framework - http://isa-specs.readthedocs.io/en/latest/isajson.html

- “The open source ISA framework and tools help to manage an increasingly diverse set of life science, environmental and biomedical experiments that employing one or a combination of technologies.
- The ISA model consists of three core entities to capture experimental metadata: Investigation, Study, Assay
- Built around the 'Investigation' (the project context), 'Study' (a unit of research) and 'Assay' (analytical measurement) data model and serializations (tabular, JSON and RDF), the ISA framework helps you to provide rich description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) so that the resulting data and discoveries are reproducible and reusable.
- An Investigation contains all the information needed to understand the overall goals and means used in an experiment; experimental steps (or sequences of events) are described in a Study and Assay . For each Investigation there may be one or more Study associated with it; for each Stud there may be one or more Assay.

Research Graph - http://researchgraph.org

- “Research Graph is an open collaborative project that builds the capability for connecting researchers, publications, research grants and research datasets (data in research). We are building a toolbox that includes the graph metadata model, open source tools that transform the connections to a graph database, and cloud-based mash-up services that connect the graph databases across multiple platforms."

- "Research Graph model was initially formed as part of the RDA Research Data Switchboard project as a graph modelling method to transform ANDS RIF-CS metadata records to nodes and relationships in a Neo4j graph database.
- Since March 2014, RD-Switchboard is lead by the DDRI Working Group in Research Data Alliance. The working group had input from a number of international partners such as AND, Dryad, CERN, da|ra, Elsevier, figshare, NCI, RMIT University and the University of Sydney."
- See also: https://rd-alliance.org/groups/data-description-registry-interoperability.html

Data Catalog Vocabulary (DCAT) - https://www.w3.org/TR/vocab-dcat/

- W3C spec.
- See also https://github.com/biocaddie/WG3-MetadataSpecifications
- "DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.
- By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation."

HCLS Dataset Description - https://www.w3.org/TR/hcls-dataset/

- W3C Interest Group Note
- Structured dataset description in RDF
- For Life Sciences, but generally
- No hard-wired serialization requirements (but download links/formats for actual data files)

ResourceSync - http://www.openarchives.org/rs/1.1/resourcesync

- Open Archive Initative (OAI), ANSI/NISO Z39.99-2017
- Intended to replace OAI-PMH
- More of a change synchronization protocol, but it has packaged bitstreams

BioSchemas - http://bioschemas.org/specifications/

- Community-driven proposed extensions for http://schema.org/ for Life Sciences (but progressing to be for any academic Research)
- BioSchemas Dataset and DataCatalog specs
- Mainly aimed as schema.org annotations within HTML web pages - no packaging

Keep Content-addressable storage for biomedical research

- https://dev.arvados.org/projects/arvados/wiki/Keep
- Has concept of a manifest

- Also collections which have mutable (like ipns?) and immutable (content addressed ids, like ipfs)

Cr8it/CrateIt

- ownCloud plugin to allow researchers to package up ownCloud-based content, adding metadata etc.
- Data packaging spec/aspect superseded by DataCrate
- https://eresearchau.files.wordpress.com/2014/07/eresau2014_submission_30.pdf
- https://eresearchau.files.wordpress.com/2013/08/eresau2013_submission_57.pdf
- https://eresearchau.files.wordpress.com/2013/08/eresau2013_submission_57.pdf
- https://github.com/digitalbridge/crateit/tree/develop
- https://github.com/IntersectAustralia/cr8it_doc
    - Active fork of OwnCloud plugin by Newcastle.edu.au:
        - https://github.com/IntersectAustralia/owncloud
    - Active fork of OwnCloud plugin by IntersectAustralia - CloudStor Collections:
        - https://github.com/IntersectAustralia/CloudStor/blob/master/collections/docs/deployment-guide.md
        - See also https://github.com/IntersectAustralia/dc21-doc (referenced here http://ptsefton.com/2017/10/19/datacrate.htm)

Metadata Schema Mapping/Crosswalk attempts

- https://docs.google.com/spreadsheets/d/1mO5nHNXpImT2w_M9VlhUJ0OeMp0x2Zq5OwiiTPXJ-M8/edit#gid=2086891330
    - Spreadsheet comparing metadata elements used in Frictionless Data packages with DCAT, CKAN, DublinCore, DataCite
- https://docs.google.com/spreadsheets/d/1XTrJgcb0OEdqSrsrfU-38BztBKm59TLDTkZ33MLViSI/edit#gid=1554679326
    - Complementing the UKRDDS metadata profile mapping document with RI data catalogues and EOSC requirements
- https://docs.google.com/spreadsheets/d/1Il2jELWZa5N-qzDA8MF0LMh_6b8fIksrCbqdyVYhF68/edit#gid=0
    - Appendix III - NIH BD2K BioCADDIE DataMed DATS v2.1 mapped to other models file (working document)
- https://docs.google.com/spreadsheets/d/1XzrZxFIuG3TS9RU8vACoUjAvaADLmI_FrIk7O3BEkxY/edit#gid=0
    - Dataset cross-walk
- https://docs.google.com/spreadsheets/d/1mjatKZKdhp_tFm6xnYJFpBgPLMNDdAue9FGy-oKFBYk/edit#gid=1554679326
    - UKRDDS metadata profile mapping document
- https://docs.google.com/spreadsheets/d/1dtHpbp5cVaooVdqhvDjLHKM5Y8IfC-iRSU6OA6BLSUg/edit#gid=1692609231
    - EOSCpilot 6.2 T2.2
- CodeMeta softeware metadata crosswalk (https://codemeta.github.io)
- DataONE metadata crosswalk

    - DataONE maps metadata from multiple metadata standards, including EML, Dublin Core, DataCite, Dryad, METS, ISO 19115, Mercury, and others
    - See details of crosswalk here: http://indexer-documentation.readthedocs.io/en/latest/

Related Discussions

- https://github.com/codeforscience/data-packaging
    - Git repo from CodeForScience for "Discussion of data formats and packaging techniques for sharing and reproducibility"
- https://github.com/ipfs/archive-format/issues/7
    - Git repo from IPFS for discussing an archival format for the IPFS p2p network
- https://github.com/frictionlessdata/specs/issues/110
    - Discussion of json-ld in the context of Frictionless Data data packages (also relevant to DAT and IPFS)
- https://en.wikipedia.org/wiki/Enhanced_publication#Re-production_and_assessment_of_scientific_experiments
    - Good references/bibliography to scholarly work in this area
- https://www.oclc.org/research/publications/library/2000/lavoie-oais.html
    - OAIS Reference Model - talks about data objects in terms of Submission Information Package, Archival Information Package, Dissemination Information Package
- https://osf.io/upwdj/
    - Container Strategies for Data and Software Preservation Workshop Videos
- Cameron Neylon on researcher perspective/needs:
    - http://cameronneylon.net/blog/packaging-data-the-core-problem-in-general-data-sharing/ (Packaging Data: The core problem in general data sharing?)
    - http://cameronneylon.net/blog/as-a-researcher-im-a-bit-bloody-fed-up-with-data-management/ (As a researcher…I'm a bit bloody fed up with Data Management)
- Wikidata and the provenance of factual statements: "Citation Needed: Digital Provenance in the era of Post-Truth Politics" (2017)
    - Wikidata is attempting to become to go o Open Science repository for statements of scientific facts. It has a pragmatic and road-tested data model for statements, citations, references, etc.
- Open Knowledge Maps: https://openknowledgemaps.org/index.php#search
    - Good example of how the value of research data increases exponentially after semantic markup.
- Wikipedia article: Comparison of research networking tools and research profiling systems
    - "Research networking tools (RN tools) serve as knowledge management systems for the research enterprise. RN tools connect institution-level/enterprise systems, national research networks, publicly available research data (e.g., grants and publications), and

    restricted/proprietary data by harvesting information from disparate sources into compiled expertise profiles for faculty, investigators, scholars, clinicians,

- OpenRIF - http://www.openrif.org
    - the Open Research Information Framework, is an open source community devoted to developing and promoting interoperable and extensible semantic infrastructure for scholarship.
    - OpenRIF community members currently work on various semantic infrastructure components:
        - SciENcv, the federal researcher profile and biosketch infrastructure
        - PARDI, the NIH Portfolio Analysis and Reporting Data Infrastructure, for research impact and evaluation
        - Attribution for many types of contributions across a wide spectrum of scholarship
        - VIVO-ISF for representing people, works and the relationships of people and works eagle-i for representing research resources, and the relationship of those resources to people and their works
            - VIVO - http://www.vivoweb.org
            - VIVO creates an integrated record of the scholarly work of your organization
            - VIVO creates a connected, integrated record of the scholarly work of your institution, ready for reporting, visualization, and analysis
            - VIVO is developing world standards for big data representation of scholarly work, providing a single shared vocabulary for data regarding scholarship
            - VIVO is creating next-generation tools for discovering who is doing what in scholarship – tools used in grant development, team-building, mentoring, and social network analysis.
            - http://vivoweb.org/sites/vivoweb.org/files/VIVO_flyer.pdf
- https://gitter.im/ResearchObject/ResearchObject
    - Public gitter chatroom of the ResearchObject project, but lots of general discussion of packaging and workflow issues
- http://ropensci.github.io/reproducibility-guide/
- https://hackmd.io/CwZgnOCMDs0LQEMAckBMdSsnMAjaIcADKgKYBmArGOSACbRJhA==#
    - Discussion in DataPackages/Frictionless Data community about defining a common API/Interface. Interesting perspectives on the role of data packages (nobody cares about them - they just need to do their job and not get in the way)
- https://discuss.okfn.org/t/w3c-csv-for-the-web-how-does-it-relate-to-data-packages/1715/5
    - Very interesting discussion on the practical reasons Tabular DataPackages and w3c CSVW spec diverged
    - Touches on the fact that w3c specs are online/webplatform first & the issues with this
- https://twitter.com/ashep_15/status/940676398116757506

- https://github.com/datproject/discussions/issues/33
    - Discussion of where tabular data spec fits with Dat peer-to-peer project