FAIR Scientific Data Management:
Connection to Digital Archiving of Historical Records

presented at:

Digital Archival Group
Minnesota Historical Society

Bob Hanson 2019.12.02

Presentation Goals

- To make the case that the task of FAIR scientific data management has strong parallels with the already extensively-developed area of historical digital archiving and preservation.
- To present a possible design perspective for our IUPAC project.
- To point out, from a digital archival perspective, what this project IS.
- To point out what this project is NOT.

“FAIR” Data Management

In 2016, the ‘FAIR Guiding Principles for scientific data management and stewardship’ were published in *Scientific Data*. The authors intended to provide guidelines to improve the **findability**, **accessibility**, **interoperability**, and **reuse** of digital assets. The principles emphasize machine-actionability.

The principles refer to three types of entities:

data

metadata

infrastructure

“**F**AIR” Data Management

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

“F**A**IR” Data Management

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorization.

“FAIR” Data Management

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing

“FAI**R**” Data Management

Reusable

The ultimate goal of FAIR data management is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Digital Archiving

Archival Science as a field is interested in preservation of artifacts in the form of an archive in such a way as to make them accessible to future scholars and/or the public.

Digital archiving (or *digital preservation)* specifically deals with electronic preservation of information relating to physical or virtual artifacts.

Digital archiving is practiced by libraries, such as the Library of Congress and the Bodliean Library, archives such as the United States National Archives, and historical societies, such as the Minnesota Historical Society.

Digital Archiving -- A Fonds

A *fonds*, as defined by the Society of American Archivists, is:

The entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator.

Or, from Wikipedia:

a group of documents that share the same origin and that have occurred naturally as an outgrowth of the daily workings of an agency, individual, or organization.

Digital Archiving -- A Fonds

It is not that we are saying that the data relating to a scientific publication or research endeavor is an actual fonds; it is that the same issues that relate to the archival of a fonds relate also to our task:

● A fonds relates to a specific individual or organization.	● A scientific paper reports the work of a research group or collaboration relating to a specific topic.
● A fonds involves a collection of related objects.	● The paper and its associated supporting information comprise a collection of digital objects.
● Archival of a fonds (these days) involves digital curation.	● Currently we have no systematic FAIR curation of the data associated with a research paper in the area of chemistry.

Digital Archiving -- Finding Aids

The field of archival science in relation to a fonds is well advanced, with substantial well-developed digital tools and specifications for the production and delivery of metadata that allow the finding of the actual objects of a digital collection.

In particular, metadata in the form of *finding aids* are critical. In relation to “FAIR”, finding aids represent the “F”; their being made accessible via digitization and cataloging amounts to the “A”.

An example of a digital finding aid is on the next slide, one of over 3000 finding aids in the collection at the Minnesota Historical Society.

Digital Archiving -- Finding Aids

One of over 3000 finding aids developed by digital archivists at the Minnesota Historical Society.

Structured, descriptive, cleanly presented -- What are we actually looking at here?

http://www2.mnhs.org/library/findaids/00008.xml

Digital Archiving -- Finding Aids

This finding aid is actually (just?) an XML document with an associated style sheet.

It is in the form of an "EAD" -- an Encoded Archival Description *(see EAD discussion)*

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="webead.xsl"?>
<!DOCTYPE ead PUBLIC "+//ISBN 1-931666-00-8//DTD ead.dtd (Encoded Archival Description (EAD) Version 2002)//EN" "ead.dtd">
<ead audience="external" relatedencoding="USMARC">
    <eadheader findaidstatus="edited-full-draft" scriptencoding="iso15924" dateencoding="iso8601"
        countryencoding="iso3166-1" repositoryencoding="iso15511" langencoding="iso639-2">
        <eadid countrycode="us" mainagencycode="MnHi">00008</eadid>
        <filedesc>
            <titlestmt>
                <titleproper>AMERICAN CHEMICAL SOCIETY: </titleproper>
                <subtitle>An Inventory of Its Records at the Minnesota Historical Society</subtitle>
                <sponsor>National Historical Publications and Records Commission.</sponsor>
            </titlestmt>
            <publicationstmt>

            <abstract label="Abstract:">Organizational records of the Section include minutes,
                reports, membership data, financial records, and correspondence, as well as
                historical information and official documents such as the charter, articles of
                incorporation, and bylaws. Also included are files concerning meetings, conferences,
                and symposia, particularly the series of undergraduate symposia held annually at
                various colleges and universities throughout Minnesota.</abstract>
            <physdesc label="Quantity:">5.2 cubic feet (5 boxes, and 1 oversize folder in a partial
                box).</physdesc>
            <physloc label="Location:">See <ref target="a9">Detailed Description</ref> section for
                shelf locations.</physloc>
        </did>
        <bioghist>
            <head id="a2" altrender="history">HISTORICAL NOTE </head>
            <p>The Minnesota Section of the American Chemical Society was chartered in 1906 and
                hosted its first national meeting in 1910. In 1965, topical groups were organized to
                support specific interest areas such as inorganic, analytical, and polymer
```

http://www2.mnhs.org/library/findaids/00008.xml

Digital Archiving -- Interoperativity

The “I” in “FAIR” stands for *interoperable*. One interesting aspect of the EAD format is that it provides mappings to other systems. For example, in this case, we see the entry:

```
<archdesc relatedencoding="MARC" type="inventory" level="collection">
```

which says that the XML tags in this document will provide MARC21 syntax equivalents. This allows the reading program to interpret fields using alternative schemas.

Digital Archiving, Finding Aids, and FAIRSpec

It’s important to understand that any finding aid for a fonds is a highly curated document. An archivist may have worked months preparing the EAD describing the contents of a fonds. The reason the process is so time consuming is that each fonds is unique and heterogeneous. There is typically no other object like it in existence, and it may contain anything from a scribbling on a piece of paper to a fully developed manuscript, a set of photographs or a stamp collection -- just about anything.

Our job is easier. We are talking about structured digital information -- chemical identifiers and structure files, experimental procedures, spectral data files, and other results of analyses. A relatively trivial case in the realm of digital archiving.

Use Cases -- Publication

An important use case is the publication of articles relating to research that involves spectroscopic analysis. For example, papers in the area of synthetic organic chemistry, or natural products chemistry.

This case can be conceptualized in terms of a fonds and its finding aids:

The *fonds* in our case might be the publication itself.

The *digital objects* of the fonds could in principle be just the two files of that publication -- the manuscript and its supporting information.

Use Cases -- Publication

For instance, in relation to the paper, *Total Synthesis of Pericoannosin A*, Daniel Lücke, Yannick LinneKatharina, and HempelMarkus Kalesse, Org. Lett. 2018, 20, 15, 4475-4477 we find, along with the paper, a 39-page PDF supporting information document.

Table of Contents **page**

Use Cases -- Publication

One approach to our task would be to extend the EAD description to provide a third publication document -- the finding aid for this publication.

This XML document would describe in some level of detail what is in the document -- structure drawing, molecular formula and molecular mass, experimental procedure, 1H-NMR, 13C-NMR spectral images and analysis, MS analysis, optical rotation, Rf, and melting point.

Note that the finding aid, as proposed here, would not include that data itself. It would simply *indicate what sorts of data are present in the collection.*

Use Cases -- Publication

However, we can do better -- by providing both a specification for a FAIRSpec finding aid as well as an IUPAC recommendation for the form of the digital data itself as separately discoverable and retrievable *research objects*. For example, we might propose using the BagIt format developed and maintained at the Library of Congress.

Note that this would allow for *distributed* data. We are not saying that the publisher would necessarily maintain anything more than the paper itself and a link to the finding aid. The actual spectra might be held at Zenodo, DRYAD, or an individual institution's data repository. (It is quite possible that even the supporting information would not even be part of the publication DOI, but this is debatable.)

Use Cases -- Electronic Laboratory Notebooks

Another use case involves ELNs such as Chemotion. When a researcher uses an electronic laboratory notebook, that framework organizes and ... TODO

The IUPAC FAIRSpec Project

Our task then, is two-fold:

1. Design a standard for format for the digital finding aids and associated metadata to accompany chemistry-related publications (or any other form of public distribution of results) that involve spectroscopic data.
2. Propose a digital format for the actual data pointed to by those finding aids and associated metadata.

The IUPAC FAIRSpec Project

Our task is NOT:

1. to require any specific data format (JCAMP-DX, Bruker, ChemDraw, SDF, NMReData, XML, PDF, etc.);
2. to require any specific place where that data will be found;
3. to actually implement any of this.

Of course, we could certainly *recommend* one or another format, particularly if it is well described by standards, and we could *specify* that certain format specifications be followed, such as a FAIRSpec extension of EAD for metadata or BagIt for actual data.

The IUPAC FAIRSpec Project

(aside)

One of the interesting aspects of NASA's Earth Science Data System EOSDIS metadata description is that it describes six different *data processing levels*, from raw instrument data to refined analysis.

Something like this would be valuable for FAIRSpec, differentiating among full FID+parameters, transformed spectra, scalable images, and crude images, for example. Or analysis description [1H-NMR (400 MHz, C6D6) δ 5.45 (bs, 1H), 5.35 – 5.30 (m, 1H),...] *vs.* a J-coupling matrix. Or an actual spectrum *vs.* a simulation.

The IUPAC FAIRSpec Project - Implementation

While implementation is not our focus, obviously we would be working closely during the development of our standards with people interested in implementation, and I would expect there to be some prototypes developed along with this project.

I suggest that it is important in terms of implementation to allow for data to be delivered in meaningful units -- that is, not just one giant 300-GB package of spectral data, but rather as individual spectra or spectral analysis components.

I could imagine that, ultimately, the “supporting document PDF” for a publication in some areas disappears entirely, replaced by just the finding aid, which could be manifested in far more useful and interesting ways using XML style sheets or other means -- including as a PDF, as we have currently.