Transcript of Interview with Bob Hanson – February 3, 2021

Interview by Jeff Lang (Assistant Director, Platform Development, ACS Publications)

Bob Hanson is Professor of Chemistry at St. Olaf College and Principal Developer of Jmol, an open source viewer of chemical structures in 3D. He is Task Group Chair of the IUPAC FAIRSpec project, which is creating standards for sharing spectroscopic data. This interview had been edited for clarity.

Jeff Lang: Bob, let's start with the basics. What is FAIR?

Bob Hanson: In 2016, the "FAIR Guiding Principles for scientific data management and stewardship" were published in *Scientific Data* (2016) and elaborated more fully by go-fair.org. The authors intended to provide guidelines to improve the **findability**, **accessibility**, **interoperability**, and **reuse** of digital assets. The principles emphasize machine-actionability (GO FAIR, 2020), as much as just personal use by humans. By the way, I interpret this as FAIR (data management), not (FAIR data) management. Which is to say that this isn't really about the data or the structure of the data. It's what is associated with the data (the metadata) and the way the metadata are managed that make for FAIR or "unFAIR".

There's a great YouTube animation created by NYU Health Science Libraries of a hypothetical discussion between a scientist who has published a paper and an oncologist who is interested in reusing the original data in the area of oncology [https://www.youtube.com/watch?v=N2zK3sAtr-4] (2012). The oncologist asks the author about their data and whether she can have a copy of it. You really have to watch this, because it will give you a great sense of why we're doing this and why this is just a nightmare right now in terms of people trying to reuse data that authors have published. The most interesting thing about it, I think, is that the author, though perhaps reticent, is trying to be helpful. They're not saying, "No, you can't have my data." They're saying, "Well, I did everything I thought I was supposed to do. Why is there a problem here?" "Seven months later" the oncologist is saying, "OK, I have the data now, and I still don't know how to use it." It really makes the point effectively that FAIR is not just "delivering the data." FAIR means making the data available and reusable.

JL: Now, let's move onto FAIRSpec, how is it different from the FAIR standards?

BH: Well, FAIRSpec is our nickname for a project of the International Union of Pure and Applied Chemistry – the "world authority on chemical nomenclature and terminology." It's a project started in March of 2020 titled *Development of a Standard for FAIR Data Management of Spectroscopic Data*. It's a little niche here – chemistry and spectroscopy. Nonetheless, it's a very important part of our field and particularly important for publishing. The project's stated objective is *"to apply FAIR data principles to spectroscopic data management in the field of chemistry building on IUPAC's extensive expertise in this area. The project will develop standards for the production and dissemination of digital data objects that contain enough spectral data and metadata that they can be (a) findable through semantic searches on the web, (b) available through standard interfaces, (c) interoperable and transferable between systems, and (d) readable and reusable over time, both for humans and machines."* (IUPAC, 2020)

Okay, so basically we intend to design a set of standards for this particular area of chemistry and this particular need in the area of spectroscopy. The project is not about implementation -- IUPAC doesn't do implementation. But obviously we want to work closely with the people who would be implementing and will be trying to do some prototyping ourselves along the way. The key is that, for example, we want to make sure that a 300 gigabyte package of data does not have to be downloaded off the web just to find out that you are not interested in it. One of the visions that I have is that the supporting document PDF for publication would be completely replaced by something like a finding aid that would be much more flexible and much more valuable than a simple PDF. In fact, the finding aid could be used to recreate customized supporting information on the fly, if it's done right.

JL: The concept of a finding aid is from library science and archiving. How does that relate to Supporting Information?

BH: I'm drawing a parallel between what we're doing in the area of scientific data management to the area of archival science, which is a field interested in the preservation of artifacts in the form of an archive in such a way as to make them available to future scholars and/or the public. Sure sounds like FAIR data management to me. Digital archiving specifically deals with electronic preservation of information relating to physical or virtual artifacts. These digital archiving practices have been widely used by the Library of Congress, the Bodleian Library, and archives such as the United States National Archives. I became familiar with this through discussions with the Minnesota Historical Society.

One of the unique concepts of archiving is this idea of a *fonds,* defined by the Society of American Archivists as *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 (SAA Dictionary, n.d.).* Or, as Wikipedia defines it, *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 (Wikipedia, n.d.).* So, what is a paper or publication in science? It's the summation of the daily workings of a group or individual on a very specific topic and associated with a set of data. I think there's a great connection. I think there's a lot we can learn from the digital archival community in relation to FAIR spectroscopy or FAIR data management.

Here's the parallel that I'm trying to make: I'm not going to say that what we have is a fonds. But a fonds does relate to a specific individual or organization, and a scientific paper reports the work of a research group or collaboration relating to a specific topic. A fonds is a collection of related objects. Well, the paper and its associated supporting information comprise a collection of digital objects, too. In fact, it is the relationship between all these objects that is often the most interesting aspect of a collection. An archive of fonds involves digital curation. But in chemistry right now we have no systematic curation of the data at all. So I'm making the case that people who are interested in scientific FAIR data management could learn a lot from digital archivists, since they have been doing this for a long time.

One of the things that I've found particularly interesting in discussions with digital archivists is the idea that when a box of historical records comes into a library or a museum and it needs to be catalogued, you get what you get. You can't demand that the data is in a certain format. You can't transform it into different formats. It's a book, it's a set of papers, it's a bunch of photographs, it's whatever it is, digital or otherwise. The task of curation is to produce a *finding aid* -- a description of the contents of that box that allows a researcher to make a query and find the description of this box along with enough information associated with it, so that the they know right off the bat if this is something that is valuable to them or not. So they don't have to hunt down the actual box and search through it.

JL: So the finding aid helps people to know where to look for the material which may still be interoperable or not and reusable or not, but it helps you to understand whether to spend your time on it.

BH: Right. It's the finding aid and that's the critical thing here.

I want to tell you a quick story here. Back in 2006, I got interested in Willard Gibbs, a famous guy in thermodynamics who was at Yale at the turn of the 20th century. I contacted Yale's Sterling Library by email and asked, "Do you have anything about Gibbs?" And they said, "Well, it seems we have a box somewhere in our archives, and it says 'Gibbs' on it, but we don't have a finding aid for it, so I can't tell you what's in it. Do you want me to go get the box and find out what's in there?" And I said, "No! I'll be there on Thursday!" I just got on a plane and went to Yale, because I wanted to see what was in the box for myself. It turned out that inside the box was a bound volume of all of Gibbs's lectures written out verbatim for a full year – three times a week for a full year. They were just a goldmine for me. If I had had a finding aid, it wouldn't have been hardly as much fun, but certainly would have been easier..

In a digital archive, sometimes all you have is a 200 MB zip file. The last thing you want to do is force a researcher to download and search through it themselves just to discover that there's nothing there for them. This is the critical connecting point, I think, between these two fields. The field of digital archival science is well advanced, and they have digital tools with specifications for the production and delivery of the metadata that allow finding actual objects. They call this the digital finding aid and it's really the F in FAIR. And it also has to do with the A in FAIR, because finding aids can be presented within a framework that allows access to them and searchability of them.

Here's an example: When you search for American Chemical Society at the Minnesota Historical Society, you get this beautiful website that talks about the American Chemical Society – Minnesota Section. If you look behind the page, you discover that it's just an XML file styled to be presented as a web page. It uses this great system called "Encoded Archival Description" (EAD) that was produced by the Library of Congress and collaborators some years ago for the production of digital finding aids.

The point is that digital archivists are way ahead of us. They have already figured out how to do what we're trying to do. My idea is that we might be able to apply this. It's important to understand that any finding aid is a highly curated document. An archivist may have spent months preparing the EAD

describing the documents. The reason the process is so time consuming is that each fonds is so heterogeneous. It may contain anything from us 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 a lot easier because we're 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.

JL: How is FAIRSpec different from supporting information that's already available with published articles?

BH: This is a really good question. So, 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 often have supporting information files. Maybe a 50-100 page PDF with a Table of Contents. So, why isn't it FAIR? First, although the supporting information PDF is findable – it has a DOI associated with it (or at least its associated article does) -- we have no idea what chemical compounds are involved. A Table of Contents doesn't cut it. In terms of interoperability, it's a zero. Even if we have a reasonably good Table of Contents, how would we ever know if there is something that we really want to look at? I routinely pull down supporting information and dig through it, trying to try to find something that interests me. Half the time it's not there. All my time is wasted. In addition, how could I ever pull this "data" into a workflow other than just reading through it? And what about reuse? Again, a zero. This isn't even the data itself. It's a facsimile, designed to support the claims in the article. Reuse is about making the actual data available in a way that could lead to unknown, unpredicted uses at the time of publication.

JL: Supporting information, that's not the same as making data findable or FAIR?

BH: Right. So, I'm the author and I think I'm being FAIR. All my information is being stored in Zenodo. Question, what's the difference between Zenodo and a flash drive? Answer, Zenodo is bigger. So, here I'm making the distinction between a *digital entity* and a *digital object*. A data collection that has no curation, no finding aid, no metadata, is just a data dump. It's potentially of some use, but it's nowhere close to FAIR. The box of historical records comes into the library as a bunch of (presumably) related entities. It is curation that turns those entities into valuable objects. Same for scientific data. The idea is that a digital entity is anything, but a digital object is part of a metadata structure that connects it to other data, provides a context, and produces value. And the key there is good curation.

I think the point that we really want to make here is this idea of digital entity vs. digital object. Just having the data is a first step, but it's not at all FAIR because we don't know anything about the data, it's just a bunch of bits. A data collection starts as a set of digital entities; our project's job is to create a system by which curators can process those bits into a meaningfully connected set of digital objects.

JL: And is that curation process happening anywhere right now?

BH: Well, it is certainly happening in some fields. Crystallographers determine 3D molecular structures and deposit those structures in databases that catalogue them and make them findable via a web portal. Structural biology has this huge set of data that is accessible, findable, and standardized in the form of the Protein Data Bank. So, yeah, there's a lot of curation that's being done -- just not in chemistry. All we have are these PDF supplemental information files. We don't have the data in any kind of structured way that could be usable. It's time to get with the program!

Here's our problem: The picture of a spectral graph by itself is useless unless you know what chemical compound is associated with it. Now, a supporting information PDF right now is generally organized by compounds. They'll show a little picture of the compound and then they'll describe the experimental procedure. Then they'll have a bunch of pictures of spectra following, so you have that physical proximity. Since this is the paragraph about this particular compound, this must be the spectrum of that compound. But, it's all visual.

JL: So, how will this support reproducibility in science?

BH: Well, I would say reproducibility is one thing, but it's not really what we're interested in here. What we're really interested in is reuse. Scientists finding use for data - the raw data – that may have nothing at all to do with the original use. Educational uses of actual data, comparison studies, investigations related to spectroscopic data itself – that sort of thing. FAIRSpec is about opening doors that we don't even know exist yet.

JL: So what will authors need to do differently to make this happen?

BH: The first thing that authors can do is to make sure that their actual spectroscopic data is not just on some flash drive somewhere. It needs to go into a repository such as Zenodo or the California Digital Library. And it needs some up-front curation. At the very least, identifying the chemical compound that relates to a specific spectrum. But, alas, we don't have a standard for that – yet!

JL: When FAIRSpec is something that they can use, what will be the benefit to the authors for the extra time it takes to make these connections?

BH: We want scientists demanding that their data be available, not dreading the task of making it so. I think if you put it the right way to PIs – that a little extra effort by a graduate student will make their life (at Ph.D. time) and the graduate student's work (all along) more efficient, they will listen. Then they need to demand that their data be allowed to be FAIR and be rewarded for that. That's where institutions can help – at tenure time, for example.

JL: So, what can libraries and publishers do to make this happen?

BH: Right now, I would say author and institutional education. We want to get the word out that the data itself is important. Anything less than putting your data up at a repository for investigation by others is

insufficient. Even that is insufficient, if the metadata is not there to make sense of it. There's no doubt in my mind that the culture is there, but researchers just don't know how to do the FAIR thing right now.

JL: The FAIRSpec project is looking for collaborators and Bob is eager to talk with anyone who'd like to participate. He can be reached at hansonr@stolaf.edu.

References

GO FAIR. (2020). *FAIR Principals*. https://www.go-fair.org/fair-principles/

IUPAC. (2020). *Project Details - IUPAC | International Union of Pure and Applied Chemistry*. https://iupac.org/projects/project-details/?project_nr=2019-031-1-024

NYU Health Sciences Library. (2012, December 19). *Data Sharing and Management Snafu in 3 Short Acts* [Video]. YouTube. https://www.youtube.com/watch?v=N2zK3sAtr-4

SAA Dictionary. (n.d.). Fonds. Retrieved on February 5, 2021 from https://dictionary.archivists.org/entry/fonds.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). doi:10.1038/sdata.2016.18

Wikipedia. (n.d.). Fonds. Retrieved on February 5, 2021 from https://en.wikipedia.org/wiki/Fonds