Interview with Arnold Rots

Arnold Rots is an Archive Astrophysicist in the High Energy Astrophysics Division at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.

June 24, 2013

Interviewers: Crystal Sanchez, Isabel Meyer, and Claire Eckert

Please start off by talking about your involvement with time based media art and digital preservation.

I don't have a background in media art; by training, I'm an astronomer. But for the last 10 or 20 years or so, I have been involved in astronomical archives, as well as file standards and time standards. The time standards we deal with have more to do with the absolute time measurement of astronomical observations in relation to where you are at the moment you actually detect a signal, which gets into all kinds of relativistic effects.

The standards work has to do largely with a file standard that was developed in the late 1970s in astronomy, when the world was a bit different. People realized it was difficult to take our digital images—which we did have in those days—and transfer them from one computer to another. Partly that was because in those days, there was no standard word length or even byte length among computers. The standard we developed has come to be known as FITS (Flexible Image Transport System), and it has expanded in many ways. One of the more interesting applications is that the Vatican Library these days is trying to archive its holdings digitally in the FITS format. I'm involved in writing a standard for time coordinates.

I've also been involved with the Virtual Observatory, which is an international effort to make the content of astronomical archives available in a homogeneous and uniform way, so you don't have to know the exact quirks of each individual archive to get data from them.

For the last 15 years or so, I have been the archive scientist for the data archive of the Chandra X-Ray Observatory, which is one of NASA's great observatories.

Please talk a little more about your work with creating standards, and the benefits you have seen from that work.

The benefits are quite obvious, but the work is very much like herding cats. Trying to get people to agree on a standard is very hard.

One fortunate aspect of FITS is that the file format prescribes that metadata and data are contained in the same file. In general, of course, in archive repositories, we tend to keep most of the metadata in databases, so the archive is searchable and users can do all these wonderful things with querying. But keeping metadata with the data means that if disaster strikes, you can always reconstruct the whole structure; I think that's important.

The FITS standard is now about 30 years old. It was started purely to allow people to transfer images on 9-track tape, but it evolved into an archival standard. The fact that it is almost globally accepted in the astronomical community certainly has made it much easier to achieve interoperability between the various archives.

However, there is one important issue that has never been properly settled. The FITS standard defines the syntax of the metadata and data structures, of course, but it does not define the semantics of the metadata. That means people in various places have been calling the same thing by many different names, which is definitely an obstacle to interoperability. Various sub-communities have developed their own conventions on the semantics they use, and tying those things together has been a major effort. For example, we have been working on the time standard for about six or seven years, and it is only now coming to fruition. It takes a long time to complete standards!

These efforts always start with a focus on conveying the contents of the data—how they were taken, where they were taken, what their properties are, and so on. Fairly quickly, however, one comes to realize that you should not be arrogant about pretending to know what metadata are really required, because it's impossible to predict how the data will be used in the future. We know that because today, when we pick up data from 20, 30, or 40 years ago, we get very annoyed that people in those days did not record everything they should have recorded. The bottom line is that you collect as much metadata as you can, with as much accuracy as you can get, because that's usually fairly easy to do at the time you create your data object, and you never know how people will want to use it and for what reasons in the future. For us, that means metadata in two areas in particular:

- *Coordinates.* We need to know exactly what the coordinates are in the sky of the images we pull together. We need to know exactly when they were taken—what the time and spatial resolution of the data are, as well as the exact wavelength response of the system. Polarization enters into it as well. You need a complete coordinate description of the data.

- *Accessory or environmental data.* These depend on what kind of observatory is used. For terrestrial observatories, we want to know the local temperature, humidity,

> wind speed, etc. For space observation, we want to know the exact location. Then of course there are all the various voltages and conditions in the entire system.

In addition, you have curation metadata—what program was this a part of; who was the principal investigator; who produced this particular file; what was the processing history (e.g. if this is a data product derived from an earlier, lower-level product, you want to know how that came about); etc.

What aspects of the data standards you have worked on—or the process of building them—would be applicable to other fields such as the arts, or any field that is responsible for large collections of digital files? I think that metadata like location might not be as relevant to the fine arts as to astronomy.

I actually do think location is important. It's now quite common for JPG images to have GPS coordinates for where the picture was taken. That may or may not be relevant for what is represented by the image, but it is certainly important for historical and curatorial records.

I can imagine some of those accessory metadata are of interest too, depending on what kind of art you are talking about—the characteristics of the digitization system, the recording equipment, and so on. If we are talking about images, you want to know exactly what the spectral response was, the spatial size of your pixels, your resolution. You need to know your sampling frequency, your time resolution—those kinds of things.

Within your own professional community, I'm sure it is extremely important to consult with other institutions and with colleagues on these kinds of matters. They have insights and experience that might turn out to be not only very useful for your own applications, but also for extrapolating into the future and for uniformity among standards—which goes to those issues of exchangeability and interoperability.

For artwork, a big part of the picture is not just the data, but presenting it to the public the right way. Are there similar issues in the data you are working with, in the sense of it being recorded in a way where in the future you either have to emulate the original environment, or recreate it in order to properly access that information?

Reproducibility of results plays a big role in our situation. That's the idea that if someone publishes something, claims a result, and draws all kinds of conclusions, someone else should be able to reproduce the same result from the raw data. If it is not reproducible, then presumably there is something wrong with the initial claim.

In principle, that is an important part of the scientific enterprise, but there are certain subtleties there that do not always make it very practical. We tend to reprocess our

observations every few years, because our instrumental calibrations have improved, our software algorithms have improved, and so on. So if someone looks at an older publication, those results will have been derived from an earlier version of the data. In principle, we keep all those versions around; but we have yet to receive our first request for an older version, because by default we make the newest version available. It's actually even more complicated than that, because people tend to forget that as well as using an older version of the data, those results may have been derived using an older version of the software. So if you really want to do proper reproducibility, you should get an older version of the software too. And even that's not even the end of it, because that older software ran on a computer that presumably was running on an older version of its operating system. So you see where this is going; it becomes in most cases totally impractical to replicate all of that.

That is one side of the reproducibility issue. The other side is purely preservation. Yes, we have lots of data that are properly backed up; but everybody keeps their fingers crossed that we don't actually have to go back to those tapes and read them, because we don't really know if the tapes are still readable. [In terms of storage at least], we do have an advantage in that even though observatories and missions are usually leading-edge in terms of the volume of the data they produce at the time they are designed, it takes a while to develop them. So by the time they are operating, that amount of disk space is usually not an issue anymore. So we usually can keep all of our data on spinning disks.

That goes to a slightly different issue. In terms of trusted preservation, you would want to have your digital objects on some kind of distributed archive system. We do that too; we mirror our archive in two places, which are about 10 miles apart—which is maybe not quite sufficient, but will do for now. Sophisticated replicating repository systems are really a must for trusted repositories, I think.

Also, I expect that for most digital repositories of any kind, if you define a particular digital object (whether that is a single file or a package does not particularly matter) that you want to preserve for a long time, you should assign a unique identifier to it, which plays the role of a URI—not a URL, because you don't know where this digital object will be living ten years from now. You need a URI or some other identifier that will remain translatable to any actual location in perpetuity, so the object can be found and retrieved. One of the reasons we do that is that we are now involved in creating links between the professional literature and the archival data. To do this, we basically connect these kinds of identifiers to articles in our digital library, so readers who browse the literature can immediately link to the archived data that form the basis of an article. And of course, it works the other way around: if people are browsing the data archive, they can immediately find the articles that were published based on a particular data set.

I glossed over the issue of deteriorating media, but that is an obvious problem. Digital standards evolve, and they may not be the same 20 years from now. For example, if you are using JPG or GIF or whatever for images, you probably will want to upgrade to better standards as they become available in the future. So you need a very clear and well-planned migration policy for how you will handle these changes in emerging standards.

How did you convince the scientific community that it was in their interest to develop standards? I'm sure scientists, like artists, all have different opinions.

I have two thoughts on that, maybe three.

First, as I said, FITS was originally developed as an image transport system. That's much less of an issue now, but back in the 1970s and early 1980s, there was no uniformity among computer architectures. You had word lengths of 16 bits, 18 bits, 60 bits, 32 bits, even 12 bits in certain places. t was almost impossible to move images from one place to another. That not only came into play when a colleague asked, "could I have your image, please?" but also when you moved from one computer to another. You couldn't take your own data along with you! So people realized standards were important for moving from one environment to another.

The second reason scientists accept these standards is that government agencies like NASA for decades have required that the missions they fund have data management plans in place: what are you going to do with your data, how are you going to preserve it, how are you going to make it available? In the space community and in the radio astronomy community, the understanding is that data will be proprietary for something like a year or a year-and-a-half, during which time the person who had the bright idea to make these observations in the first place will have a chance to analyze them, write a paper, and get credit. But after that period, the data become public property that everyone has access to. That approach has been widely accepted—least of all in the optical astronomical community; but even there, its time will come. So NASA says "as part of the contract, you are required to provide us with X, Y, and Z so we can have a complete record." [You could perhaps also] have that kind of requirement on contracts with artists.

(By the way, people also realize that if they see something they think is new, it can be very helpful to go back to photographic plates from 100 years ago and see what things looked like then. The Harvard College Observatory here has half a million or so glass photographic plates spanning almost a century that they are now trying to digitize at great expense.)

I do have to admit that we have not been terribly successful in getting people to provide us with links between the papers they publish and the data they used. People don't really

want to bother with putting all of that extra information in there. However, we have discovered that articles that have links to data are cited more often than the ones that don't, so we are now using the self-interest argument: "Look here, the more you give us in terms of metadata and links and what-have-you, the wider your work will be publicized. So it's in your own interest to provide all this stuff."

Could you discuss the distinction between actual standards and institutional best practices in your field?

It's a fine line. The problem with standards is that it takes a long time to establish them. You need to be extremely careful because you are going to lay down the rules for a long time to come. You cannot afford to change your standards every other year. Standards have to remain in force for extended periods; otherwise, they are totally useless.

Also, you need backward compatibility. You do want data objects created under previous versions of standards to remain valid under new versions.

Conventions, best practices, guidelines—these things tend to develop among communities, and are often useful as precursors to standards. In our case, that is precisely what happened with the standards that were eventually developed for coordinate systems. They were not part of the original [FITS] standards, but people clearly and quickly realized that they were needed, so they started developing conventions. Fortunately, the community was not too splintered on that.

There are other areas where you don't particularly need standards, but where guidelines and conventions and best practices are sufficient. That is certainly the case where you have certain things that are relevant to a small sub-section of the community. For example, I suppose in the artistic community, standards will be segmented by media. I can't image that there would not be some issues that would be best handled through conventions within specific media-segment sub-communities, and need not be imposed as standards across the whole field. But I don't know. Anyway, you also have to be open to the possibility that conventions will one day become part of a standard, if they are experienced as something that is useful for a large part of the community.

Are there any other areas where standards you know about—storage, acquisition, etc.—might be relevant for our purposes?

In terms of types of standards, there are quite a few. You go with ISO standards wherever you can. OAIS is extremely useful. Within the Smithsonian on an earlier go-around, we have been looking very seriously at TRAC. There is the U.K. digital conservation project, which is doing very nice work on standards. I think it is important to lay down your standards in

policy documents, to make sure that people are aware of them as early as possible. Trying to impose standards after the fact is virtually impossible.

What training and resources do you rely on for some of the challenges you were talking about, such as migration and awareness of technological changes? How do you stay abreast of the current status of things?

I entered this area a long time ago, when things were not terribly well-organized and were very much "seat-of-your-pants." In many ways, we in the astronomical community are very lucky to be a relatively small but very international community. The discussion among software-oriented and preservation-oriented astronomers is fairly lively. People get their information from their colleagues.

Another issue here is that astronomers got into software development quite early on, so most astronomers—at least the ones of my age and younger—are usually fairly good programmers. I don't mean to say they are professional programmers, but they have an easier time communicating with software engineers and can talk on that wavelength. That helps them to absorb the information that comes from the software field, and gives them a better sense of what is feasible and what isn't, even when there will be quite strong debates on the best way to move forward.

Also, in astronomy there is relatively little off-the-shelf software that we use. We generally develop it ourselves—"ourselves" meaning not just the astronomical staff, but the software engineers we hire. That makes for quite a different environment from the average digitization shop. In addition, practically all of our data today are born-digital, which is different from most other fields.

Do you see any aspects of the work you do in digital preservation or digital archiving that are challenging right now, and that need more research or attention?

Interoperability is a big thing; also development of persistent identifiers; and most definitely, standards on trusted digital repositories. A number of systems around now are quite good implementations of the distributed repositories model. I'm not so sure the Cloud adds much to that, but that's another story.

Did we forget to ask you anything?

One thing that occurred to me was the issue of compression versus lossless preservation. If you can preserve your digital objects in a lossless way, that is definitely to be preferred. If you use a lossy JPG, for instance, there's no way that later on, when better and more accurate standards become available, you will be able to retrieve the original image. FITS

images are by definition lossless—even though we do use compression, the compression techniques are lossless techniques. It makes sense to use compression, but not at the cost of losing information.

Do you work with video?

No, but our images can be turned into video sequences. We store every pixel in its original value, and our images in principle are n-dimensional—in time, wavelength, frequency, polarization, anything you have. The solar X-ray astronomers have beautiful video clips produced from the original lossless images of the sun. (See Solar X-Ray group website, http://xrt.cfa.harvard.edu. See also the helioviewer at http://jhelioviewer.org, although I think that was produced by ESA.)