



Address 100 Barr Harbor Drive  
PO Box C700  
W. Conshohocken, PA  
19428-2959 | USA

Phone 610.832.9500  
Fax 610.832.9555  
e-mail [service@astm.org](mailto:service@astm.org)  
Web [www.astm.org](http://www.astm.org)

---

**Committee E13 on MOLECULAR SPECTROSCOPY AND CHROMATOGRAPHY**

**Minutes for E13.15 Subcommittee Working Group**

10:35 am – 12:30 pm EDT

May 26, 2006

Virtual Meeting

- I Introductions and Welcome:** Gary Kramer, E13.15 Chair called the meeting to order at 10:35 am EST. A poll was taken, and with no objections, the meeting was recorded.
- II Attendees:**
- |                        |                          |
|------------------------|--------------------------|
| Mark Bean, GSK         | Richard Larsen, JASCO    |
| Michael Boruta, ACD    | Peter Linstrom, NIST     |
| David Farrusseng, CNRS | Alex Mutin, Shimadzu     |
| Maren Fiege, Waters    | Alexander Roth, NIST     |
| Ronny Jopp, NIST       | Burkhard Schäfer, BSSN   |
| Joe Koury, ASTM        | Wolfgang Winter, Agilent |
| Gary Kramer, NIST      |                          |
- III Capturing minutes from the core group:** Burkhard Schäfer talked about a new technique used to record the minutes of the core group virtual meeting. When there was something to be documented, someone typed it into the notes section of the Live Meeting window. At the end of the meeting, the minutes were almost complete, having evolved during the meeting itself. The consensus was that this was easily done with a small group, but would not be manageable for a large group such as today's meeting.
- IV Next virtual meeting:** The next virtual working group meeting will be held on June 30, 10:30 am - 12:30 pm EDT.
- V Approval of minutes:** Minutes from the April 21 meeting were posted. Maren was surprised to see her name as a potential speaker at a meeting in Rostock. She had not been planning to go, and noted that the abstract submission deadline had passed in March. Burkhard replied that since he knows the organizers, we could probably still get in. Since this topic was discussed during the April 21 meeting, that portion of the minutes will not be changed. David Farrusseng noted that his name was misspelled on the 2<sup>nd</sup> page. That correction will be made. A motion was made and seconded to approve the minutes as amended. The motion carried.
- VI Changes at Sourceforge.** SourceForge is in the process of changing their procedures, with an impact on the CVS environment and the rsync service. Hostnames and path names are going to change. Because of this, Burkhard has been unable to post the PittCon presentations. However, he has written some instructions for working in the new environment, and will send these to the developer list. The changes should ultimately make SourceForge faster and more reliable.
- VII AnIML PR.**
- For PittCon, we didn't get approval for our proposal for a symposium, but we can get a workshop. The difference is that for a symposium, Pittcon pays for speaker travel and only 4 speakers are allowed, while a workshop can have 4 or more speakers, but no travel support. The consensus was that we should ask for a ½ day workshop. The deadline for submission is August 1. PittCon 2007 will be in Chicago February 25-March 2, 2007.
  - Anand Mudambi and Joe Solsky invited E13.15 to participate in the National Environmental Monitoring Conference, the week of August 28 in Arlington, VA. Gary Kramer and Dale O'Neill will make presentations.
  - Gary Kramer received an invitation for an speaker on AnIML at a Laboratory Informatics Conference, similar to IQPC, being held in Philadelphia, August 1-2. He never heard anything more, so assumes we missed this chance for a presentation.

- Burkhard and Gary participated in a conference call with Joe Koury and Richard Wilhelm of ASTM. Gary answered a number of questions, and Richard turned the answers questions into a short article, for *ASTM Standardization News*. The idea is to make other ASTM committees aware of our work, to encourage collaboration with other groups working on XML specifications.
- There was discussion about other publications which would be useful for E13.15 to try similar articles. *LC/GC* and *Spectroscopy* are sister publications, so it might be possible to get similar stories in both. The details could be similar, but the *LC/GC* article could use examples from chromatography, while the *Spectroscopy* article could use examples from spectroscopy.
- **Action Item:** Wolfgang Winter will contact Advanstar to see who we should approach to see about a story on AnIML for *Pharmaceutical Technology* and perhaps other publications as well.
- A Life Science Automation Conference, sponsored by CELISCA, is being held in Rostock, Germany, September 14-15, 2006. The deadline for submissions has passed, and it is unclear whether there is still a chance for E13.15 to participate.
- David Farrusseng is writing a report for major chemical companies - Shell, Exxon, etc, and would like to get two pages from ASTM article to add to that report. The report needs to be submitted by June 26. He will submit to Gary Kramer for review. Burkhard will send the *JALA* article.
- The Euro CombiCat Conference will be held Apr 22-25, 2007 in Bari, Italy. There will be room for 1-2 presentations on AnIML. *Applied Catalysis* will publish the presentations. The abstract deadline has not yet been set, but will probably be sometime in November. David will send the first flyer to the developer list. The second flyer will fix the schedules.
- Burkard Schäfer gave an invited talk at Roche's global QA/QC Workshop last Tuesday. Eighty people from all Roche sites attended. This was an internal meeting, but they traditionally invite one guest speaker. Burkhard noted that they already knew a lot about AnIML, and are very interested.

**VIII Core Developer Meetings.** The core group has had some meetings, and would like to discuss some questions with the main group.

- In AnIML, references are used as a mechanism to link related experiment steps. For example, a spectrum might occur at a specific point in a chromatogram (a data point) or at a specific point in time (a value). Two possible ways to link the spectrum to the chromatogram are to say that the spectrum occurred at  $t=30$  sec, or at data point #278. An index oriented approach allows some simplicity, since if there are multiple independent vectors, only a single index is needed. However, a disadvantage is that you can only point to data points that exist. Allowing arbitrary values gives more flexibility, since the references don't need to snap to a grid. However, allowing values makes the schema larger and more complex, and also makes validation more difficult. Annotations are normally aligned on a pixel, not a data point, so the need for values rather than data seems to be required, despite the desire for simplicity. If we did both, that would solve the shortcomings of either solution alone. However, the implementation would then be more complicated, since both methods would need to be supported. It was noted that referencing by value seems more inclusive. It does everything we want it to do; whereas if we pick index, we'll be constrained. We will defer indexing to be considered for version 2, after seeing whether reference by value is sufficient.

**Motion: To use values instead of indexes because we may want to annotate areas in a graph where there are no data points.** Second. The motion passed.

- There was discussion about how to use name/value pairs when only a single value is involved, such as melting points. The schema is looking for a single value, but it could be a range. It was noted that we already allow a minimum and maximum value for both melting point and boiling point. If you only have one value, then the maximum should be used. Maximum is required, minimum is optional.
- The topic of uncertainty was discussed. It was noted that people handle uncertainty in values in different ways. In addition to uncertainty, we may need to consider bias in measurements as well. These topics are common to consider within NIST, but are not widely considered in many other areas. In general, analytical chemistry tends to be very weak in handling these. The issue of how statistics enter into this area was also raised. ThermoML, which reports results instead of data, has the capability to handle uncertainty. It was also noted that in many cases, these values come through user interaction, and not from the instrumentation. We should probably create an optional container for this, so every report out of an instrument doesn't need to have it. But if AnIML is adopted within the chemometrics community, it will need to have it.
- Another discussion centered on the best way to relate sequences of data. Do we need to store more than one sequence in an AnIML file? In general, we have considered storing individual measurements, chromatograms, and spectra, which may be part of a specific sequence. However, there may be a quality calculation which is based on analyzing a different sequence of measurements, rather than a single measurement. How can one you specify which

data belongs together? We have defined roles for standards and blanks. But in some cases, an analyst may pull a sample from a production batch, so the sample may be both real sample, and a test sample. Multiple test procedures may require system suitability samples, blanks, and standards before running a real sample within a production process. Some people do this on the instrument itself, while other people use their lab notebook. Each run could be a separate discrete sample. Even though we can't support a complete GLP (GxP) model, AnIML should at least support this kind of sequence, or grouping. It was noted that in clinical environments, accession numbers are used to track samples through processes. For example, in blood tests, a serum separation may create multiples samples all derived from a single sample. There is a logical connection between things that belong together, but there is not a universal way to do this. It was noted that such tracking might already be possible within AnIML, but that this use case might be compelling enough to consider specifying a way to handle this in a well-defined way. It was also noted that for 21CFR11 compliance, we need to be able to store all of original data. If structural information is included, like sequence id or accession number, we can't drop it. We need to include the information, but we don't need to enforce its inclusion where not needed.

The questions are how far do we need to go in support, and what are the consequences to the data model? The consensus was that there are a number of valid reasons for grouping measurements, or providing some sort of reference to external data, within AnIML data. A related discussion concerned whether AnIML files should contain multiple data, such as chromatograms, creating large files, or whether multiple data should be spread across multiple files. There is no requirement either way. Storing separate files would require replication of some metadata, but AnIML will support either method. We may also want to do multiple runs, but only save a single run. In research, for example, one may do a series of runs to optimize, but only want to save the last one. Alternately, in a QC environment, you probably need to keep every iteration. This becomes an enterprise decision, a business rule that needs to be implemented outside of the core, and enforced by an enterprise extension. We don't want to force people to store things they don't really need. Some vendors now require GLP data stored by everyone, but in AnIML, if you need it, you can use it, but if you don't need it, you don't have to use it.

Some type of tagging is needed to designate groups of measurements, but everyone does the grouping differently. One possibility is to consider attaching something like "luggage tags" to an experiment step(s) or sample(s) to identify them as part of a group, and assign an arbitrary identifier to the group. The data system or a specific application would need to understand the context of the identifier, but AnIML doesn't really care what it is. The analogy to a luggage tag is appropriate, since although it usually is for an individual, a tour group might have a tag to identify individuals with the group. Such a tag could be used throughout AnIML for an external reference, such as to an enterprise extension. It may not mean anything to anyone else. The assumption is that the information on the luggage tag would be defined by the user application. AnIML just defines the place to put it. It should be allowed, but not required, on both experiment step and sample. However, we also need to make sure people don't get the idea that they could put other parameters into these tags. If they do, we'll have incompatible flavors of AnIML. To help prevent this from happening, we need to write rules to specify what we have in mind into the NDRs. Then we need to trust the implementers. The question was raised whether these tags should be restricted in length, as one method to prevent their misuse. The consensus was that they should not be limited. If you set a limit, someone will find a legitimate reason to do more.

**Motion:** We will add tracking tags to the AnimL core, which can be added to experiment steps and samples, zero to many, they serve the purpose to group related samples and related experimental steps together, and also tying them to external references. Tracking tags may not contain any attributes that are technique specific. They just serve for grouping and identification purposes, and are optional. Their semantics are defined by the data system or external application the data file is produced by. They have an xml type of token or string and should be constrained appropriately so that they cannot easily be abused. Defining the restrictions will be delegated to the developer group.

**Discussion:** It was noted that xml type token excludes cr/lf

**Motion Seconded. The Motion passed.**

- Auxiliary techniques will be discussed in the next core group meeting.

## IX Miscellaneous:

- XML Review: When we start looking examples of XML, what is the best way? Should someone with XMLSpy act as presenter? Not all have XMLSpy, and XML Notepad doesn't always work. A suggestion was made to try using Internet Explorer, since it can display XML files.
- An issue has been raised as a result of Robert Lancashire's work on a generic Java based viewer. He had problems using AnIML data as a stream. It is difficult to relate sample information to the data in a single pass because of the

use of references. A pure hierarchical tree, with no references, would make it easier to do this. In the current structure, all sample information is at the top of the file, and sample roles are not known. It is not until more of the file is consumed that one discovers the role, or roles, of a specific sample. To process the file in a single pass, one would need to keep the sample information in memory, and refer to the in-memory storage when needed. It is possible to modify the schema to store the role within the sample definition. If you need a new role for the same sample, the sample would need to be redefined. This could cause problems by forcing duplicate data storage. It was noted that this is a classical problem of data parsing.

The concern was expressed that someone is actually using AnIML, and is frustrated. However, what he is asking for is a major disruptive change, which would reduce the functionality of AnIML. It was noted that a year ago, the committee discussed whether the file should be completely flat, with references, or completely hierarchical. The decision was made to implement as a tree, but also to allow references. There is a problem, perhaps, because we have allowed both.

A question was also raised about sample compound relationships - a sample can contain many compounds, and a compound can exist in more than one sample. Sample-compound is a many-to-many relationship, and can't be expressed in AnIML. As an example, consider a link between compound and chromatographic peak. A peak can contain more than one compound, and a compound can exist in more than one peak. It was noted that this is post-analysis piece of data, and perhaps outside the scope of the design requirements. On the other hand, in order to support legacy systems, with annotation capabilities, we might need to have some support. We have talked about annotation in the past, and the question arises as to what actually needs to be linked to the sample.

As a final note, the developers decided in the previous discussion that the hierarchical tree structure would be used, but that some situations required the use of references. Therefore, both are allowed, and the schema will remain unchanged.

**X For next meeting:** NDR for UV/Vis and NDR/Schema for UnitsML. Also note that UnitsML OASIS Technical Committee will have a preliminary conference call, with an official meeting coming soon.

**XI Adjournment:** The meeting was adjourned at 12:30 pm EDT.

Minutes taken by David Martinsen, via recorded session