

AAS/CDS/IOP meeting (August 31 and September 1) agenda

Venue

Observatoire astronomique de Strasbourg, 11 rue de l'Universite, 67000 Strasbourg
03 68 85 24 10

Meeting room basement of the Grande Coupole building (the one with the large dome)

Participants

AAS: Greg Schwarz and Gus Muench

IOP: Graham McCann and Julian Norman (IOP)

CDS: Pierre Ocvirk (PO), Mark Allen (MA), Francoise Genova (FG), Emmanuelle PERRET (EP), Sylvain Guehenneux (SG), Patricia Vannier (PV), Thomas Boch (TB), Gilles Landais (GL), Soizick Lesteven (SL)

Arrivals

Gus will be in Strasbourg (Hotel Gutenberg) Saturday night.

Greg will be in Strasbourg (Hotel Mercure Strasbourg Centre Petite France) Sunday afternoon.

Graham and Julian arrive Monday morning and be at CDS by 12:30.

Departures

Greg, Gus, Graham, and Julian have a flight to Amsterdam Tuesday evening at 18:35.
A taxi will collect the visitors at 15.45 at the main gate.

Topics

- 1) Short discussion of the ongoing changes at AAS given the framework of the Journals taskforce decisions and how they might benefit CDS.
- 2) Further integration of AAS data products into the CDS data service.
 - a) Common CDS problems that could be addressed during data preparation at the AAS. [Pierre has already sent around a document outlining the 4 issues that consume the most time at CDS.](#) PO: we at CDS think this should really be the core of the meeting, we want to go through the points raised in detail and see how these problems can be solved.

["Missing" MRTs:](#) AAS-IOP uses a row number cut to determine when a table should be standardized. AAS-IOP understands the different scientific needs at CDS.

Question: could AAS apply the simple rule:

- standardize the table with the coords as well even if small
- if a standardized table has links to another table, standardized the linked table as well.

These 2 rules should cover a fair fraction of the cases.

How much tables/work would this add to the AAS processing?

=> **ACTION [CDS] 15** to quantify the number of small tables collected by CDS not MRT's by AAS, and send on to AAS. If reasonable, **ACTION [AAS] 14**: in future references, look for associated tables which are linked from MRTs or containing positions from MRTs' objects.

Tables standardized by CDS: use of the xml files still under evaluation, they could indeed be useful. GS acknowledges that broken pure ascii tables (non-MRT) is a problem is a problem for AAS as well, not just for CDS trying to ingest them. AAS states that having access to the latex file would not always be very helpful as latex tables can be quite complex rather than being a nice simple array.

ACTION AAS 6,9: continued effort to improve pure ascii tables as well.

ACTION CDS 7,8: further evaluation of xml by GL and CDS documentalists

No coordinates: coords are often removed by authors to because they use lots of columns (the paper legacy). All agree that the electronic table could have more columns/info than the paper table. In the new aastex there may be a new command for hiding columns of a table, which would allow the author to still have the coords of objects in the table, but just not show them in the paper version, while they would appear in the ascii and standardized versions. Additionally, attempt to better communicate with the authors, with CDS in the loop.

ACTION 20,21

Truncated IDs:

ACTION CDS 11.1,11.2: CDS will supply simple workflow diagrams for external audience, as well as 1-2 "horror stories" illustrating the problems caused by bad or truncated IDs in tables, to be added to the authors guidelines. The authors should also be given incentive to improve their tables, for instance accelerated availability of their data at CDS (15 min vs 6 months).

In general, attempt at improved curation by AAS: **Action [CDS] 16**.

- b) Discussion on the usefulness of UCDs and VOTable vs MRT or another standard.
CDS will keep doing UCDs anyway, because they are useful for mining and for the documentalists ingesting the catalogs. PO suggested AAS could do less/no UCDs but focus on more MRTs. To be re-evaluate after action 15's results are known.
- c) CDS techniques/tools for handling large data input and service ingest that might be useful for the AAS as the number of articles with data increases.
PO: can not remember we've discussed that?
- d) Effectively communicating CDS discovered mistakes to the AAS for resolution.

CDS will add a keyword in catalogs metadata flagging tables which have been significantly edited wrt the paper version. AAS-IOP will be able to retrieve these flags. **ACTION [CDS] 18**

- e) Integrating CDS corrections and value add to the XML article. Is the “CDS” link under the “Related links” sufficient or needed? Future annotations layer discussion?

CDS will send a weekly feed describing new vizier tables to AAS-IOP which will allow AAS-IOP to link to vizier tables.

AAS-IOP can query simbad by reference sometime after publication to retrieve the CDS-validated list of objects found in the article.

ACTIONS [CDS] 3.1, 18,

CDS can explain how to build a “vizier widget” displaying a quick, simple summary view of the catalogs, a la simple vizier page

(<http://vizier.u-strasbg.fr/index.gml>) to be displayed on AAS-IOP website.

Part of ACTION 2.

- f) What about data that will not reside on IOP servers? What does CDS need to effectively use this information? Could CDS be an external site for complex data sets, posting prior to publication and long term identifiers?

This will certainly be the topic of an ongoing discussion.

ACTION 2

- 3) Verification tools and support for non-tabular data content. Can tools be created to check meta-data against VO standards (modified ObsCore?) for authors to use prior or during peer review? One example would be a FITS header checker for spectra.

This is ongoing work, as shown by GL’s demonstration of an interactive software module converting fits headers to ObsCore, as a part of the new submit page for associated data. The CDS Obscore construction procedure has been applied only to a very small fraction of vizier data and is still very much in alpha stage. In the future, once we have a working validator, it would make sense to share it with AAS/IOP. More news at AAS in January from GL.

ACTION [CDS] 10,19

- 4) Unified Astronomy Thesaurus and semantic enrichment of AAS content. Can CDS provide guidance and/or support? We are running a test case for facilities with the astro librarians. Can we replicate facility and object matching with similar semantic searching? GMc: Do CDS have any controlled vocabularies that AAS/IOP can use for the enrichment of content, and linking of articles to CDS landing pages. I have in mind <http://cds.u-strasbg.fr/cgi-bin/Dic-Simbad> and what up-stream use can publisher make of this that would help link records? If we identify objects in research articles from this list and use Sesame / XML Web Services, would we be able to enrich articles in a useful way; and useful to whom!?

CDS point of view: Extraction of objects from references is very hard to automate: at CDS is strongly supervised. Therefore CDS thinks the most economical way for AAS/IOP to retrieve/link list of astronomical objects is to pull them from the simbad database (query by

reference) sometime after publication. CDS will send a feed of recent refs processed by SIMBAD. **ACTION [CDS] 3.2**

Wrt the thesaurus results for an article, we at vizier would like to see how the thesaurus-returned keywords could suggest keywords for vizier catalogs. Keyword attribution in vizier will remain supervised but could the thesaurus speed up this process / help documentalists? One important aspect is that the keywords attributed by documentalists relate to the tables, not the body of the article. CDS agrees to evaluate usefulness of the thesaurus-provided keywords. IOP can also check its thesaurus results against human-attributed keywords in older vizier catalogs (using the README). **ACTION [CDS+IOP] 12**

- 5) GMc: I'd also like to discuss linking directly into Sim-Play <http://cdsweb.u-strasbg.fr/SimPlay/> and Aladin Light.

Following Aladin-lite demo by THomas Boch, CDS is willing to help AAS-IOP to add Aladin lite visualization on their website, for instance for associated fits images. **Action [AAS] 4**

- 6) GMc: Review the Astronomy Image Explorer for potential linking to CDS services. Can we take data from CDS to make links back to CDS for particular images. How easy would it be to make this into a process, and what would the benefit be?

- 7) PO: A demonstration of the ingestions of AAS tables into vizier will be made by the CDS documentalists to illustrate the process.

JULIE comments (08/18/15)

Gus and Greg: Please bring the assembly up to speed on the Sloan and AZ Data Hub projects, respectively. We want to involve CDS as one of the trusted data repositories with which to connect and/or federate for both of these projects. Currently, we are talking about infrastructure that is similar to what iPlant has at Arizona and CERN has for Zenodo. In particular, I'd like to know what the CDS take is on persistent identifiers and workflows.

GM exposed project for software + version citation

ACTION MINUTES from the two-day meeting

1. [AAS] Next meeting will be in January, 2016 at AAS meeting in Kissimmee, Florida
2. [IOP, CDS] Francoise to act as a point of contact with data citation of tables. Look to pull a catalogue summary from Vizier to display on IOPscience, once we know that Vizier have used the data. Looks like a single line entry, with option buttons to access the data. How do DOIs get built?
3. [IOP, CDS] CDS will send two different feeds, by email or ftp, and IOP to look at ways of displaying the data against the full text and elsewhere on the article page:
 1. [Pierre] Vizier Catalogue

2. [CDS] Simbad Objects
4. [AAS] look for opportunities for interactive figures that will convert static images than an author suppliers into an interactive version using Aladin light. Early action after interactive figures project. Use Aladin Lite in as a FITS viewer now, as apart of interactive figures
5. [IOP] See if semantics can pick up coordinates and automate display as a pop-up using Aladin Lite; focusses on visualisation; Easy to embed; Javascript API
6. [IOP] Number of gif images of mathematics have increased from about 10 over an issue to over 100; in abstracts and titles. This takes an hour of time a day at CDS to deal with, used to take no time at all. soizick.lesteven@astro.unistra.fr, Soizick Lesteven in contact with Chris Ingle. Fix our ASCII pipeline. Pick up with Chris Ingle about Mathjax and broken images; and ascii tables. Greg to talk to this: where +- images are dissapearing because we don't have the correct encoding.
7. [IOP] Gile wants to covert XML tables to CSV; can our typesetters do this for us or can we transform the tables automatically; would like to be able to download the tables as CSV automatically for an article. Chris Ingle to help with this. JSON, ASCII pipleine. CDS would like delimiters between ASCII columns
8. [AAS] Need a sample data set of the XML / ASCIIs/ MRT / LaTeX from the past six months for some articles to allow CDS to look at the data and work out the best format that they should use. Send on to Gille. ApJ/2013//771/110 was a complex example.
9. [IOP] We should be capturing the AAM as PDF in the processes (need to check this with Chris Ingle) and we should be able to do this with the LaTeX file too; and supply to CDS if that works out best.
10. [AAS, CDS] Create a page demonstrating good usage of FITS, post January.
11. [AAS, CDS] Make it clear to authors the sorts of processes that the data from their papers are used for; inform authors and encourage them to put more effort into data preparation. Given them a better feel of the level needed.
 1. [CDS] Simplify CDS workflow diagrams for an external audience, for January meeting
 2. [CDS] Ingestion horror stories to illustrate the effects of poor data quality
 3. [AAS] Same diagrams and information for peer review and production.
12. [IOP] Thesaurus keywords trial; quick
 1. Send list of current extractions to CDS for quick review
 2. Also do a small trial ADC list
 3. IOP looks at ADC Keywords list <http://cdsarc.u-strasbg.fr/doc/ADCkwsd.hlx> pass this list to Michael Roberts; the left-hand-side of the list
 4. This is fairly fuzzy; not much opportunities here because the data is needed at a table level, not an article level. Consider carefully whether we do this trial for CDS
13. [AAS] Change to TOCs have lost some of the links to MRTs that CDS archivists have been using. Need to be careful about consulting when we make these changes.
14. [AAS] Look for associated or object tables that are referenced, containing positions or links from MRTs.

15. [CDS] to quantify the number of small tables collected by CDS not MRT's by AAS, and send on to AAS.
16. [AAS, CDS] Curation changes; hot list of labels. Stick with the labels when possible and be consistent. Is there a class of labels where we have the most problems. Be explicit in explanations of the definitions. [Check and explicit filter systems \(AB vs Vega, Johnson, Cousins etc...\)](#). AAS see an opportunity to do better here.
17. [CDS] Update to 'prepared by AAS', replacing Greg's name for future records, in Vizier.
18. [CDS] Add a keyword into the metadata, to flag tables that have been significantly edited by CDS, so that IOP/AAS can identify tables that have been updated and can apply to the article as an update, possibly as far as an erratum. What level of change would trigger this keyword, need to be clear what will trigger the addition of a keyword to stop lots of small changed being flagged. Defer to January discussion.
19. [AAS] Using FITS checker software? January discussion.
20. [AAS] Try to meet IAU recommendations on nomenclature, working with authors, but need to understand the limits on what can be achieved and how far AAS is able to influence authors.
21. [AAS] Arrange for CDS to review the information used in author workshops. before the next workshops in January 2016. CDS may be able to participate in January, and would be keen to get involved on a longer timeframe; e.g. IAU in Vienna, 2018. [Greg] to mention to Julie.
22. [Pierre] to merge actions list with minutes list, based on Pierre's notes in the Google Doc.