eIUS: Astronomy Experience Report

In the text: Some barriers in '{{…}}' to provide more contextual information.

Interviewee profile

Senior Lecturer at a UK university's (Edinburgh) astronomy department

Time spent in research

10% of time at the moment (i.e. individual research); "I also lead a unit which helps lots of other astronomers doing research" in curating data that is used by a lot of other astronomers, in the UK and across Europe, as the basis for their research.

Research area

Research interests: cosmology; formation and evolution of galaxies and clusters of galaxies; multi-wavelength survey astronomy; astronomical data-basing;

Research question(s)

"Understanding evolution of galaxies, especially (.) [that of] the galaxies which are in clusters of galaxies (..) we have lot of evidence that (.) being in a cluster (.) alters the evolutionary history of a galaxy, but it's not understood in detail how (.) [that] actually happens – what the actual [physical] mechanisms are. So I'm interested in looking into observations which will help understand the particular evolutionary history of the galaxies which are in clusters of galaxies."

Research Lifecycle

Literature Review (channels, repositories) – Start of the research process

Astronomy is not a large discipline: the researchers "feel very lucky" that a restricted number of major journals exist, and those collaborate a lot;

Usually starts the research process using/querying an online repository called ADS (http://adswww.harvard.edu/ ; website: "The SAO/NASA Astrophysics Data System (ADS) is a Digital Library portal for researchers in Astronomy and Physics, operated by the Smithsonian Astrophysical Observatory (SAO) under a NASA grant.". It contains scans of older journal articles and a database of online versions of more recent articles. "I'd always (..) [start with] using ADS as (..) [my] route into the literature."

Data collection process

Historically "an individual astronomer had to go and make observations out at the telescope, and then analyse all the data [themselves]; (..) [they'd then] write it up into (..) journal [articles] and that'd be the end of it": Nobody else would have ready access to their data, even though it could be very useful for other research projects.

Today large areas of the sky are mapped in larger projects, observing in different bands of the spectrum .The data is loaded up into a database and is accessible online at a data centre. National and international institutions host open access data in observational archives for the community, with different data centres specialising in the curation and publication of data from different regions of the spectrum: for example, the Wide-Field Astronomy Unit in Edinburgh specialises in optical and infrared data, LEDAS in Leicester on x-ray data, Manchester on radio data, etc. Usually data is embargoed for 18 months while it is used by its owners and then becomes open access data.

A challenge is to make all these archives interoperable, so that users can perform different kinds of analysis on combinations of the various kinds of data (optical, x-ray, radio) to get "a new perspective on the objects of interest". An international initiative – the "Virtual Observatory (VO)" – is looking into this task of achieving interoperability and defining standards for describing data, using metadata registries, and access protocols. The "move into having standard means of access (.) [to] the data (..) means [that, as a researcher,] you can expend all your effort actually analysing the data not just (.) getting it [and integrating it]".

"[Currently] you have to know which of the archives (.) are likely to have any data you want to use, and then you have to visit each of them on the web [individually]. You have to learn how you access any of the data in them; (.) each of them has its own means of query[ing] (.) [its] underlying database and they will each spit out (.) data products" in a particular format, which you'd then download and manipulate them into a format you could analyse. The aim of the VO is that you should have a registry that you can query to determine which archives have relevant data, and then each archive uses standard query protocols and data export formats, so that you do not need to learn the idiosyncratic methods of each one.

Examples of archives ("I can email you afterwards and [give you] some url's"):

1. The Wide-Field Astronomy Unit here curates the WFCAM Science Archive (http://surveys.roe.ac.uk/wsa/), which houses data taken by an infrared camera on a telescope in Hawaii, and is about to start the operational phase of the VISTA Science Archive (http://www.roe.ac.uk/~nch/wfcam/), which will house data from a new survey telescope, VISTA, located in Chile. Another example of a state-of-the-archive is that of the Sloan Digital Sky Survey (http://cas.sdss.org/astrodr7/en/)

2. Regarding Virtual Observatory development, that work was being done in the UK by the AstroGrid project (see http://www.astrogrid.org) until it was axed due to the current STFC funding crisis. The international Virtual Observatory initiative is coordinated by the IVOA (see http://www.ivoa.net).

3. The data centre in Strasbourg which implements the links between its archives and papers in "Astronomy and Astrophysics" is the CDS (see http://cdsweb.u-strasbg.fr/).

Repository content (example)

For the optical and near-infrared data that is curated in Edinburgh, image analysis codes are run over the image data to generate catalogues of detected sources. These sources are described in the catalogues using basic standard attributes, which record things like "where it is, how large it is, is it elliptical, is it round, does it look as if it's

a galaxy or does it look as if it's a star, measuring the brightness" and so on. These attributes are then uploaded into the database and are the object of most analyses (not the images themselves). The image analysers are not perfect for all purposes, so, if someone wants to analyse something more unusual, the images themselves might be analysed again for that.

Data analysis

"And then you could either analyse them [data products] using some standard libraries and packages or you can actually write some code [of your own to] (.) analyse the data – and it's a mixture of the two in general. (..) I think one of the things that's distinctive about astronomy is we do lots of different things with our data. If we don't have standard libraries and tools implementing all the things we want to do (..) often we have to write our own code to do the particular things we want to do with our particular data products."

The degree to which astronomers can rely on existing code from code libraries or packages depends on the data, e.g. in space astronomy "each mission has its own code library that analyses its own data products", "you would have to learn that package and then you could use it". In the general case, for performing a particular analysis on an "image or spectrum or catalogue of objects and attributes describing them" it is usually necessary to "write your own code".

The languages and packages used also change over time, i.e. depending on which the user was trained. The interviewee uses IDL, a package for data and image analysis, while younger people seem to use Python more. "Almost all astronomy groups in Britain have IDL licenses."

For the analysis stage mainly scripting languages are used, but for the earlier stages in the data production process data centres tend to use "real" programming languages, like Java or C/C++.

{{For the future it is seen as useful to increase functionalities of repositories for early stage analysis by providers ("move more of the data analysis stage into the data centre, (..) [so they then] have to offer at least some sort of standard algorithms that they're expecting users [are] going to use on their data") because "archives are getting bigger and bigger and bigger and we're reaching a stage that an individual astronomer is unable then to download [onto their own workstation] all the data they want to analyse onto their own workstation because they will be doing analyses using millions or billions of objects and they just actually can't download all that data" over the network.}}

Currently "almost all analysis is (.) run on catalogues of objects which are extracted" from images. Each "of these objects is described actually using a list of attributes, and almost all of the analysis is [run] on these tables of attributes describing objects which have been extracted out of images. As it's tabular data it's easily loaded up into a relational database and you would nowadays expect [data centres to allow you to run] SQL queries against that. Then obviously that enables you to discover objects with any particular properties or count numbers of objects with those particular properties, things like that". These SQL queries enable filtering data in an automated way, so this can already considered being a part of the first analysis steps ("we can issue queries which build new histograms of attributes, (..) which means you can do at least the

actual basic [statistical] analysis of the data just issuing queries on the database"). At this point they enable users to do plots of histograms or scatter plots (i.e. graphical representations of data sets) which are then just displayed in a browser and the dataset as a whole does not have to be downloaded (the data stays on the data centre's server). Overall these features are still "very basic at the moment" and can not address more complex astronomy questions, which still require downloading all the data for further analyses on their own. {{"We'd like to build on that and be able then to do data-mining and sort of a large scale and statistical analyses of the data": it's clear that will become increasingly important in the future, but for that "we haven't [yet] been able to get (.) [adequate] funding".}}

Astronomy deals with observational data (not experimental data) and the "analysis is looking for correlations between different attributes of different kinds of objects" to come to an understanding of the underlying physical mechanisms at work in them.

Example: "I'm interested in these issues about of the evolutionary history of the galaxies and clusters of galaxies which means I do things like look for correlations between attributes which describe clusters and other attributes describing individual galaxies in them, and then you look for any statistical correlations and things like that. And then you also have (.) [theoretical predictions] as well", which models "how the evolutionary history of the stars and galaxies work, how they age and how they look over time, and you look into the matching your observations against the predictions of these models" to test them.

"I don't actually build any of the theoretical models".

Collaboration

The interviewee is collaborating with other researchers: this is the norm now, as projects grow larger and become more and more international.

Use of tools: email; f2f project/collaboration meetings from time to time for discussion and planning ("email and then you have collaboration meetings every now and again which everyone discusses the things that they've been doing individually and then you discuss what you're going to do next"). Access Grid is used very occasionally. Collaborative visualisation tools are not used at all, although it might be handy to have them at times.

Example: The interviewee gives feedback to colleagues/collaborators on theoretical models based on the experience he makes from using those models in analysing the data. This happens directly through interaction but also indirectly through papers ("describe things you've discovered observationally which their models aren't able to explain then that's also giving them comments on their models, but indirectly").

Dissemination

Describing work on one particular paper (exemplifying the general research process): One paper recently submitted to a journal described work from a long-term project involving the interviewee. It is about the analysis of x-ray data from an archive and started some years ago, to look for clusters of galaxies in the context of the evolution of the galaxies within them, i.e. "looking in our catalogue of clusters of galaxies and trying then to work out which are the young ones which are lying at larger distances" to get indications of galaxies in a young, early evolutionary stage.

"And there aren't a lot of them". X-ray data is particularly useful for discovering clusters of galaxies and a student developed the algorithms for that, as part of his PhD project: "And over the years we built up this catalogue of hundreds and hundreds of clusters that we've discovered in this x-ray data archive." Cross-correlations with this data and data from optical archives helped to identify the individual galaxies in the clusters – also some new observations had to be made of clusters without existing data in the optical archives. With this data set the next step was to further study the clusters of galaxies and discover 'interesting' clusters, so that further observations of those could be made. 'Interesting' in this context means young: the researchers are "actually looking back into the early history of our universe, into the early stages of cluster evolution; (.) [these] clusters are rare objects and (.) [it's] interesting" to discover the "really early ones." The paper is describing these observations and "illustrates that these projects often (.) run over many years and have different stages in them: of analysing data which are in archives, then actually making new observations, analysing those data and then using those data and to think (.) what are the (.) [further] observations which are needed and then making new observations and analysing those data and that's sort of how it all moves on."

Important journals: Nature is the most prominent one, but it only accepts certain types of paper, so: "Here in the UK we (..) [mainly] publish in the 'Monthly Notices of the RAS' [= 'Monthly Notices of The Royal Astronomical Society, MNRAS'], which is a British journal."

"In the US there's the 'Astronomical Journal' and the 'Astrophysical Journal' and [in] Europe there's 'Astronomy and Astrophysics'. Those are the main journals and almost all results [in] astronomy end up in one of those four."

Journals and online content/repositories: "If you publish anything in Nature then you have to give extra stuff that's offered online and there's descriptions of the data that you used etc. The other astronomy journals (.) don't require that, and indeed most of them (.) make [quite] a (.) distinction between the description which is in the journal and the data which underlies it, and I think actually that's a mistake".

The CDS data centre in Strasbourg (see 3. under 'Examples of archives' on p.2 of this document) has specialised in collecting and curating astronomy data. They were early definers and adopters of standards for metadata and things like that and are the leaders in that overall area, in astronomy. They collaborate with 'Astronomy and Astrophysics' providing links to the source data from the text of the online articles. {{"That's obviously needed in the long run and ought to be used everywhere I think, but it's (..) [on quite a] restricted scale operation at the moment, especially because the thing naturally needs a lot of (.) manual effort to check all the links and to get them all implemented in the journal. (..) In the long run we do need to have more links between the online literature and the databases which hold all the observational data."}}

In general more and more authors from astronomy start to put links to source data into their articles to enable the reader to query the actual underlying data.

Other important elements about/in the research:

{{A lot of research is done in collaboration with international partners, but most e-Science infrastructure projects are national – the National Grid Service (NGS) doesn't

help Canadian or German project partners. Data ends up being open access, but in the 18 month period when it is embargoed and only used by its owners secure access to archives is needed. That means they require interoperable authentication schemes at an international level.

"(.) we've had to build our own infrastructure into the virtual observatory which (.) builds in all the international inter-operability amongst astronomers and we've ended up (.) using all of our own stuff, not all of the standards in the grid world, because they don't actually meet our needs."}}

"Astronomy is mature enough that (..) astounding discoveries [are not made] every day: (..) it's largely incremental and you only make big changes over years of effort". "You might (.) be working on a particular problem over many years but you'd expect that you'd be publishing a number of papers on the way". "Astronomy is really observationally led and you can have (.) new instruments [coming online] (..) which do things (.) that you just haven't been able to do previously, and then overnight you have a great discovery just because you are using an instrument which is able then to (.) [make] observations that nobody's been able to do until yet" – and after a new instrument starts operation there tends to be a flurry of papers from it reporting new phenomena.