Highlights of Spanish Astrophysics VII, Proceedings of the X Scientific Meeting of the Spanish Astronomical Society held on July 9 - 13, 2012, in Valencia, Spain. J. C. Guirado, L.M. Lara, V. Quilis, and J. Gorgas (eds.)

Digital Science: reproducibility and visibility in Astronomy

J.E. Ruiz[1], L. Verdes-Montenegro[1], S. Sánchez[1], J.D. Santander-Vela[1], and J. Garrido[1]

[1] Instituto de Astrofísica de Andalucía – CSIC

Abstract

Most of the science done in Astronomy is digital science, from observing proposals to final publications, data and software used: each of the elements and actions involved in the overall process could be recorded in electronic format. This fact does not prevent that the final result of an experiment is still difficult to reproduce. At the same time, we have a rich infrastructure of observational data and publications. This could be used more efficiently if greater visibility of the scientific production is achieved and seamless reproducibility guaranteed, which would avoid duplication of effort and reinvention. We present the current results of the Wf4Ever project. In particular, how the use of scientific workflows as the digital characterization of the methodology may boost visibility and reproducibility of the scientific outcome, hence its discovery, re-use and a more efficient exploitation of present astronomical archives, computational infrastructures and observational facilities.

1 Introduction

One of the current challenges in Astronomy is the efficient exploitation of the huge volume of data currently available, whether generated by experiments/observations, or computed by means of numerical simulations. This efficiency is needed in order to ensure the prompt return of the big investments made in terms of facilities to obtain those data, something that clearly the traditional methods of analysis are not currently achieving. A systematic capture of the scientific process may allow researchers to create, re-use, and share the full methodology and agents of an experiment, whilst reducing effort duplication and ensuring repeatability of the discovery process, which finally leads to a complete change of paradigm in the way research is performed.

Ideally, astronomers should work only with data ready for their scientific interpretation, by means of a set of analysis tools that would cover all possible use cases. Alas, this is not

the current situation, due to a vast variety of specific needs for the analysis of the processed data, which are not covered by the most common software packages. As a result, the most widespread working methodology combines general-purpose software with specific tools developed within a single research group, and based on the extremely valuable knowledge and experience of a limited set of people. This procedure does not scale with the complexity level and the size of the data being generated by the new facilities, which double every year. Hence reducing reinvention and effort duplication in Astronomy is a must.

Traditional paper publication does not support reusable and reproducible research, hence new mechanisms are needed in order to publish knowledge and not only advertise results. In the Wf4Ever project we propose to improve the quality of science with metrics based on reproducibility and reuse, preserving decomposable thoroughly curated digital artefacts that enhances reproducibility and visibility of the experiment, as well as allowing more accurate mechanisms for credit attribution.

2 Data archives and digital libraries

The quantitative leap in volume and complexity of the next generation of astronomical archives will need analysis and data mining tasks to live closer to the data, in computing and storage distributed environments. These tasks should also be modular enough to allow customization from scientists and easily accessible to foster their dissemination among the community. Because of the size and complexity of the next generation of observational datasets, data providers will have to store huge volumes of data, but at the same time should supply on-line processing and analysis services that will provide on-the-fly generated data. These services could be used as components of internet-based workflows that capture and preserve the scientific methodology in a rich ecosystem of interconnected archives speaking a common language of web services. At that moment, processes should benefit of the same curation and preservation privileges acquired by data until now. Moreover, storing digital recipes as data generators instead of final data products is a topic of great relevance in the upcoming context of Big Data.

Reproducibility is a cornerstone in scientific method, but the process of reproducing an experiment in Astronomy can be long, tedious and not easily accessible, reliable or understandable, even to the author. It seems clear that a digital experiment that cannot be reproduced, repurposed or re-used is not as useful as it should be. Present efforts to mitigate these issues are related with Open Data policies (covered in EU Horizon2020 research programs), as well as incentivising scientists to thoroughly document their results and provide datasets behind the plots.

Existing studies have shown that the top barrier for the scientists to document their results in a reproducible way is the time required for creating documentation [10]. These practices are not only laborious and time consuming, but what is worse, they are not properly

rewarded. The obvious incentives to produce a fully documented and annotated digital experiment are: allow later re-use by the creator, sharing within the research group/collaborators, and training of e.g. new PhD students. Making those methods public has some handicaps: a) editorials do not ask for them, what is more astonishing, often do not even ask for the used data to be made available, b) lack of a proper citation methods imply sadly a risk of plagiarism.

Other studies have shown that citation rates are higher for those articles where the scientists spent some time providing links to digitally published data [7, 8]. A thoroughly annotated digital experiment will greatly boost the visibility of the scientific research, considering annotations as part of the weights in algorithms for recommendations and indexed searching, which in turn will raise the citation rates. Visibility is one of the most coveted goods in science, and it could be used as a highly significant incentive for providing well-documented digital experiments. Moreover, the possibility to link and add external URL resources would improve the visibility of any other hidden gem as a digital resource available on the Web (blogs, videos, posts, web pages, unpublished documents, etc.)

Astronomy was one of the first disciplines to benefit from the early developments of Internet and web-based technologies to enable cross-linking of resources across archives [1]. In the last year, the US Virtual Astronomical Observatory (VAO) has launched an initiative [3] to create an infrastructure supporting curation, discovery and access to VAO resources. The two main objectives of the project are to capture and describe the linkage between astronomical objects, archival datasets from surveys and catalogues, observing proposals and publications and to capture and describe as much as possible the lifecycle of the research process, thus enabling to track the provenance of both data and publications produced by researchers. Thanks to this backend server infrastructure a new ADS search prototype is being built in order to expose these links, making it easier for astronomers to explore the space of astronomical concepts and phenomena using an iterative process through an interface which exposes key relationships among them [2, 4].

Astronomy research is entirely digital, reproducibility and intellectual contributions are mostly encoded in software developments, credit attributions are not decomposable enough to map authoring on specific parts of bigger and more distributed experiments, and actual publications do not expose the complete scientific record but a story advertising the final achieved results [6]. In this context, we consider that time has come to go *beyond the PDF* as the only option to publish the scientific outcome.

3 Preserving reproducible science

Since December 2010, the AMIGA group (Analysis of the interstellar Medium of Isolated GAlaxies, IAA-CSIC) is heavily involved in the EU FP7 funded project "Wf4Ever: Advanced technologies for enhanced preservation workflow Science". Wf4Ever aims at providing the

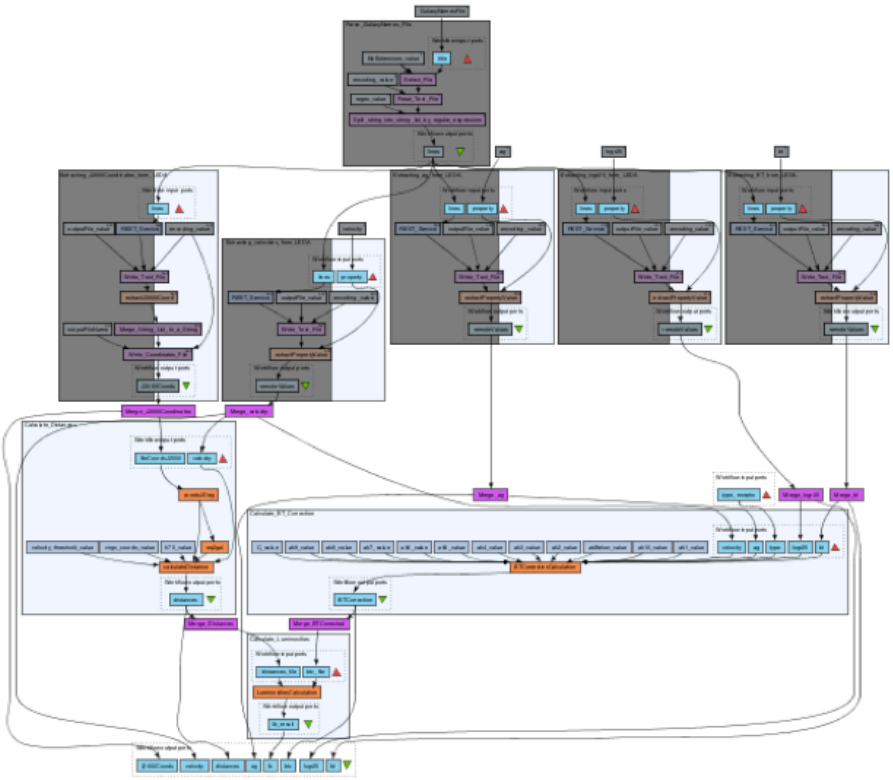

Figure 1: Screenshot of a Taverna scientific workflow during its execution

methods and tools required to ensure the long-term preservation of the scientific methodology in scalable semantic repositories in order to facilitate their discovery, access, inspection, exploitation and distribution among the community. These repositories store the experiments on "Research Objects" [5] whose main constituents are digital scientific workflows, the provenance of their executions, and links to all the related resources upon which they depend. The scientific workflows [9] provide a comprehensive view and clear scientific interpretation of the experiment (Fig. 1) as well as the automation of the method, going beyond the usual pipelines that normally end up in automated data processing.

Wf4Ever is developing models for repeatability and reproducibility, and models for workflow abstraction, to facilitate classification and indexing, comparison, and similarity detection between pairs of existing workflows and ROs in the library. Strategies for sharing and reusing workflows, ROs, their fragments and patterns, including mechanisms for personalised recommendation based on descriptions, users collective behaviour, and social information are also considered. Wf4Ever will develop a software architecture and reference implementation for the preservation of these digital artefacts, which will extend one of the most widely deployed scientific workflow sharing infrastructures (myExperiment) with preservation capabilities that consider the complexity of scientific workflows and their related objects.

At the start of this project most of the required infrastructures were either in place, or being developed with the goal of the interoperability of both data and methods through the Virtual Observatory initiative, although essentially no scientific workflows existed in Astronomy. The contribution of the AMIGA group is directly related with the development of a representative set of Golden Exemplars: a) handling of large 1D tabular datasets of physically interconnected parameters, b) extraction and characterization of sources from digital surveys and public 2D images, c) kinematical modelling of 3D velocity datacubes of galaxies. Then, subsequently assemble, pack and annotate all of the involved pieces through preserved Taverna based scientific workflows and research objects.

Automation of tasks is a pressing concern that has been successfully solved in Astronomy with scripting in different program languages and environments, depending on the specific astronomical domain of research. Consequently, the added value arising from the migration of existing scripts into workflows is not the automation of the process, but the improvement in the transparency of the experimental protocol. This allows the astronomer to precisely know how to execute the experiment, which datasets are needed and how to set-up the execution environment. This knowledge is hidden in the scripts, preventing the reproducibility of the experiment and hampering the replicability of digital science.

On the other hand, the migration of the existing experimental protocol related with user-interactive graphical software, into a more automated digital flow has made us notice the potential impact of workflows used as living tutorials. Scientists may visualize the actions performed by the workflows as they progress in their executions, allowing them to practice self-learning by the example, which expedites training and avoids reinvention. We are convinced that digital libraries of workflows and research objects will boost the use of the existing rich and underused infrastructure of data in Astronomy, and Virtual Observatory archives in particular, since they will provide these missing living tutorials on how to use them. Big storage and computing distributed infrastructures (HPC, Grid, Cloud) will be also exploited due to easing the application of the developed methodology to larger samples requiring higher computational power. Astronomers in the AMIGA group are already benefiting of some of the workflows developed in the Wf4Ever project, in order to re-use them as templates for similar experiments (e.g. extraction of luminosity profiles in radioastronomical images).

Preservation is deeply related with conservation tasks performed on the archived ROs when facing a potential reconstruction in order to bring a RO back to life. At this stage, provided annotations pertaining the recovering process play a crucial role. Information on authoring and credit attribution is useful to achieve long sought citation rates, but most important this entails responsibility. We consider of great relevance the possibility to register the tuples user-annotation, as well as the provenance indicators who, when and why, in order to know who to blame and asked for specific issues related to his contribution. This practice should in principle modify the existing citation system, enabling credit attribution to specific parts of the experiment as well as different roles in the contribution.

4 Conclussions

In this coming era of data intensive and digital science, it is increasingly important to be able to seamlessly move between scientific results, the original data and the processes used to produce them. In addition, scientific research requires that we are able to establish the provenance of data sources and processes operating on this data, so that research may be repeated, variations on the research be carried out, and new research on existing datasets be enabled. While the Virtual Observatory has provided us with the standards and protocols needed to model and exchange astronomical datasets, it does not provide the means to curate and preserve the actual processes crunching those data as well as the experimental protocol followed, moreover information discovery could still be greatly improved. The AMIGA group try to cover those gaps in the framework of the Wf4Ever project, developing models and tools that will allow publishing of decomposable reproducible experiments, as well as new metrics for their quality assessment, credit attribution and discovery.

Acknowledgments

The research reported in this paper is supported by the EU Wf4Ever project (270129) funded under EU FP7 (ICT-2009.4.1) in the area of Digital Libraries and Digital Preservation. This work is also partly supported by Grant AYA2008-06181-C02 and AYA2011-30491-C02-01, co-financed by MICINN and FEDER funds, and the Junta de Andaluca (Spain) grant P08-FQM-4205.

References

[1] Accomazzi, A. 2011, FPCA-II. vol 1, p. 135

[2] Accomazzi, A. 2010, ASP Conference Series, 433, p. 273

[3] Accomazzi, A., & Dave, R. 2011, ASP Conference Series, 442, p. 415

[4] Accomazzi, A., Kurtz, M.J., Murray, S.S. 2010, Proceedings of Science, 27th IAU GA SpS5

[5] Bechhofer, S. et al. 2010, FWCS2010 Proceedings

[6] Braun, T. et al. 2010, Nature, 465, 870

[7] Henneken, E.A., Kurtz, J., Eichhorn, G., et al. 2006 Journal of Electronic Publishing, 9, 2

[8] Henneken, E.A. & Accomazzi, A. 2011, ASP, ADASS XXI Proceedings

[9] Schaaff, A. 2011, ASP, ADASS XXI Proceedings

[10] Stodden, V. 2010, NIPS 2010