Evanthia Hatziminaoglou
European Southern Observatory
ehatzimi@eso.org

Douglas R. Clark
La Sierra University
dclarck@lasierra.edu

Evanthia Hatziminaoglou and Douglas R. Clark

Data and Metadata Standardisation and Sharing and People-Powered Research: Synergies Across Disciplines

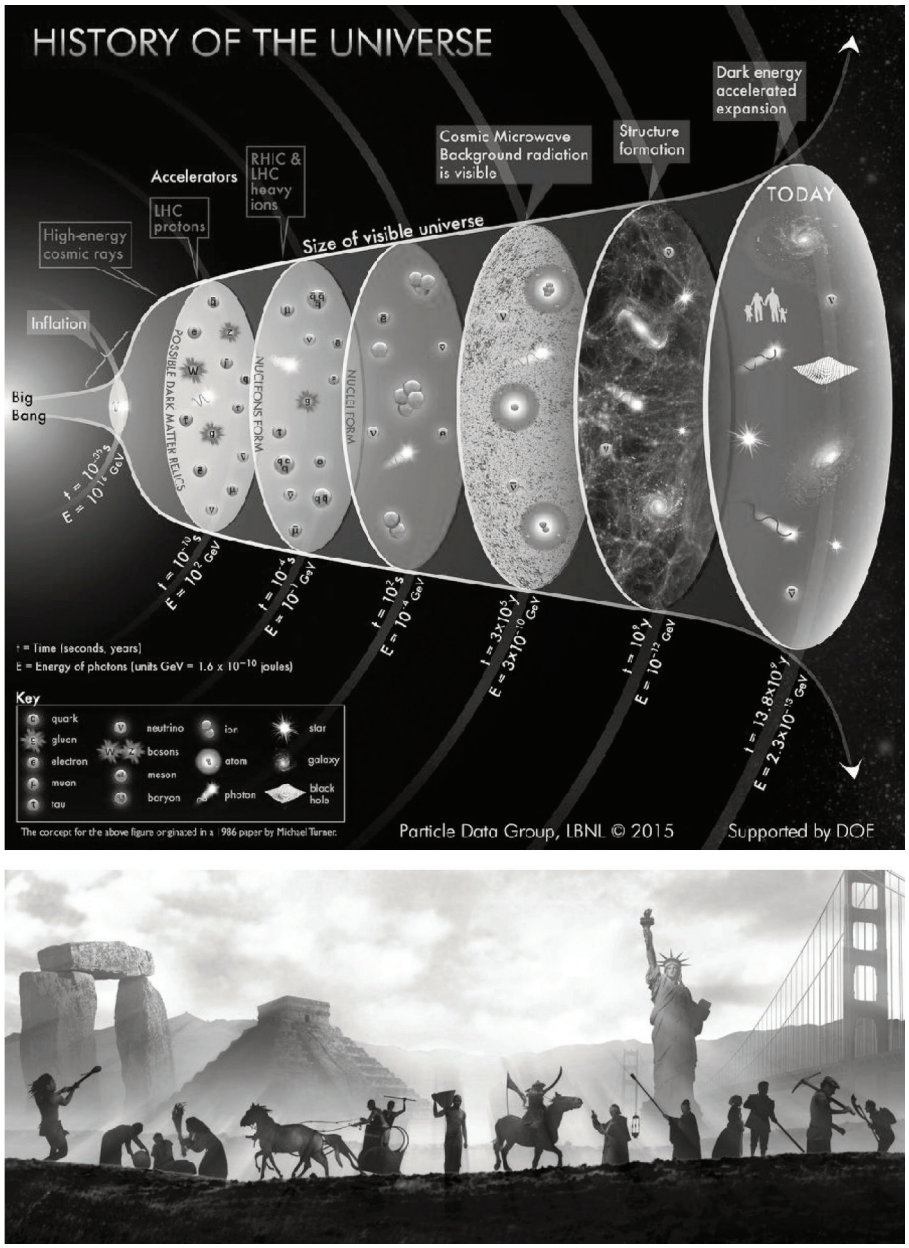
Abstract

Astronomy and archaeology are two seemingly disparate disciplines that share a common goal: the search for our origins. Other than the shared philosophical implications, this common quest also leads to a plethora of conceptual and methodological commonalities. Beyond data acquisition, data standardisation, sharing and interoperability efforts to bring together and curate data sets and other research resource, preservation of their provenance, searchability of metadata, and seamless accessibility of all those resources are only some of the challenges the two disciplines are facing in common. At the same time, with the advent of technology, both fields of study start benefiting strongly from the contribution of the general public in the form of crowdsourcing or citizen science.

Introduction

Since the 17th century, and in particular since Isaac Newton's *Principia,* all scholars know that all scientific methodologies share certain characteristics, namely systematic observations, measurements and experiments, accompanied by formulation, evaluation, and subsequent adjustment of hypotheses or theories. Nevertheless, it was not until I joined the Madaba Plain Projects ʻUmayrī excavations in the summer of 2010 that I realised how much astronomy (my field of expertise) and archaeology shared in common. The one obvious similarity, the search for our origins, is of course the main driver for both pursued via attempts to reconstruct history (of our civilisation or that of the Universe; Fig. 1). But there is much more to it than that. Simple things in common include having to work at hours outside the usual nine-to-five, as both astronomers and archaeologist tend to (Fig. 2). Data acquisition is almost exclusively based on observations supported by models, while experiments are largely beyond reach. But deeper and

Studies in the History and Archaeology of Jordan XIV:
Culture in Crisis: Flows of Peoples, Artifacts, and Ideas
Amman: Department of Antiquities, 2022



1. Top panel: A brief history of the Universe since the Big Bang (Image credits: Particle Data Group, Lawrence Berkeley National Laboratory); bottom panel: A brief history of human history (Image credits: history.co.uk).

more complex commonalities can be found in the scientific methodologies employed and their underlying implications. Parallels can be drawn, for instance, between digging and observing (*i.e.*, data acquisition; Fig. 3); cleaning object and reducing data (*i.e.*, the 'cleaning' of the raw data as taken at the telescope to produce *e.g.*, science-quality images); storage of artefacts in museums and astronomical data archives, just to

2. Top panel: Sunrise at the Roque de los Muchachos observatory La Palma, Spain (taken in June 2005); bottom panel: Sunrise at Tall al-ʻUmayrī (July 2010).

name a few. At the same time, while time is the most important factor for both, albeit at very different scales, the two disciplines also face similar challenges, namely weather and human interference. As time goes by, precipitations can destroy any man-made construction while the same phenomena will prevent observations carried out by any optical telescope. And humans are adding to the disturbance by *e.g.*, looting or

3. Top panel: Control room at the Atacama Large Millimetre/Sub-millimetre Array (ALMA) Operations Support Facility at 3000 m in the Atacama Desert, Chile, where astronomical observations with ALMA are run (August 2018); bottom panel: Digging at Tall al-ʿUmayrī (July 2010).

encroachment on one side and light pollution or unregulated commercialisation of space (*e.g.*, commercial satellite launching) on the other.

Perhaps more importantly, methodological similarities include the 'big data challenge' as well as all the issues inherent to data and metadata standardisation, data availability, accessibility, and sharing. Given their almost universal nature across disciplines, these issues are discussed in what follows from the prism of a perhaps naïve astronomer that, having invested a number of years in addressing them in her

own field, sees a huge potential for cross-disciplinary development of methodologies and subsequent applicability.

Big Data

The 'big data challenge' can be described in terms of data volumes, data rates, and the need for almost instantaneous data availability and accessibility to scientists across the globe. The challenge is not unique to a specific set of disciplines. It has spread across fields from biology (the pioneer in the domain) to genetics and medicine, to physics and astronomy, but also outside science, in business and industry (*e.g.*, transportation, logistics, automotive, manufacturing, etc).

The challenge of big data presents the different scientific fields with common issues. They include (but are not limited to) data and metadata standardisation, digital data storage, maintenance, handling and accessibility, data download tracking, proprietary versus public data, data curation, ownership, and referencing. The answers to some other related questions, however, are discipline-specific. Data combination for instance means different things to different people and while in astronomy it most commonly signifies putting together data taken by different telescopes and/or different wavelengths on the same object or region of the sky, in archaeology it could imply folding in data from other fields, such as climatology. Data storage to posterity is also a notion with different meanings, given that in astronomy we are interested in the maintenance of data already digitized at the stage of acquisition while in archaeology a lot of the data are buildings or objects that need physical space to be stored and maintained, with digitisation and subsequent storage in digital archives being a separate step that comes with its own complications.

As archaeology is going digital it is inevitably finding itself in the big data quandary. In order to avoid the proverbial reinvention of the wheel, methodologies can be borrowed from other disciplines and adapted to the needs of the field. Examples of some related efforts made within the field of astronomy and beyond, with possible relevance and applicability to archaeology are given below.

The Need for Standardisation

Astronomical data are taken by ground-based or space-borne telescopes (mainly images, spectra, and time-series) or produced by simulations. Data from different telescopes could come in different formats (especially in the early days of digital astronomy), with datasets containing enormous amount of information organised in ways that could be fully structured to fully unstructured and anything in between. Information is usually extracted from those data and stored fairly often in tabular format, but the shear amount of information creates a complex, multi-dimensional problem in terms of data storage, preservation, management, accessibility or visualiation, and documentation.

In a loosely drawn parallel, archaeological data are of two distinct categories: material data such as architectural items or objects (including artefacts) and intangible data, *i.e.*, measurements (*e.g.*, mass, material, colour, texture), direction, and/or orientation and associations (*i.e.*, positioning the material data in relation to their surroundings). The amount and diversity of the data and documentation, inherent even to the initial steps of the archaeological research, such as data acquisition, leads to similarly complex data handling issues.

Standardisation is the pillar for data preservation and seamless data sharing regardless of the discipline, as described by Whitcher Kansa, Kansa, and Schultz (2007): 'Among the primary technical and conceptual issues in sharing field data is the question of how to codify our documentation. Archaeologists generally lack consensus on standards of recording

and tend to make their own customized databases to suit the needs of their individual research agendas [...].' And even though this refers explicitly to archaeological data, the problem is anything but unique to archaeology.

Data standardisation **is** vital if we want to bring data into a common framework that allows for collaborative research, the sharing of tools and methodologies, or large-scale data mining and analytics. In astronomy this effort started in 1981 with the development of a format in which all astronomical data are stored in a structured way. The Flexible Image Transport System (FITS) is an open standard that defines a format used for data storage (Wells *et al.* 1981). One important feature is the storage of metadata in ASCII (plain text) format in the header of any FITS file, that can easily be extracted and that provides all the necessary information regarding the provenance and characterisation of the data.

Description and characterisation of metadata is equally important as standardisation of data formatting and storage. Similarly, long-term storage and accessibility as well as access to published results via dedicated platforms are essential to ensure future exploitation, repeatability of results and usage of the data beyond the original scope, by the expert communities and the public.

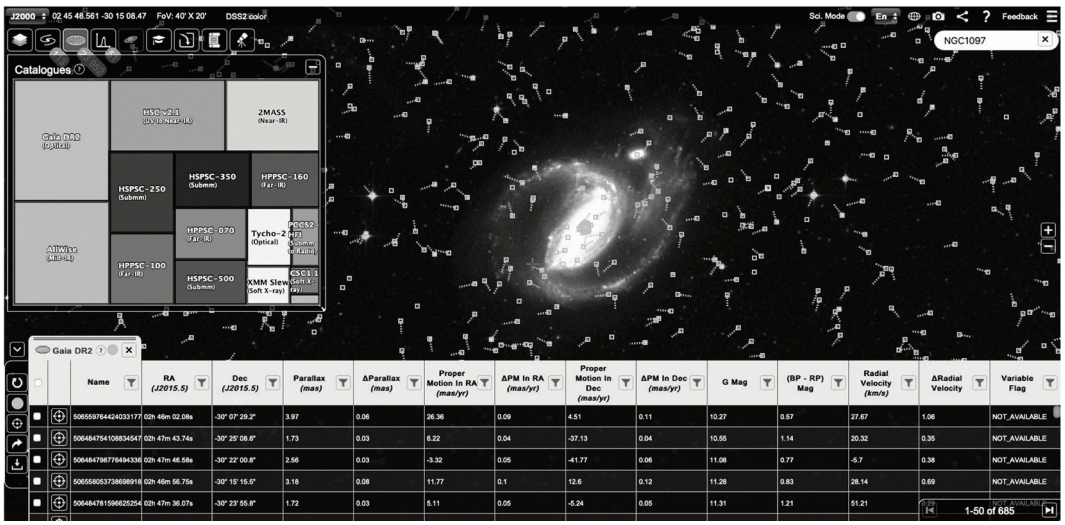
The astronomical Virtual Observatory (VO; Szalay and Gary 2001) is an international community-based initiative, an aggregation of interoperable data archives and software tools that form an environment in which original astronomical research can be conducted via the internet. The VO allows transparent and distributed access to data by developing and promoting common standards and by ensuring interoperability between the various data collections, tools, and services. It is driven by the vision that astronomical datasets and other resources should work as a seamless whole, to open up new ways of exploiting the huge amount of data continuously collected by astronomical facilities and computer simulations (Hatziminaoglou 2010).

To this aim, the International Virtual Observatory Alliance (IVOA; ivoa.net) was formed in 2002. It is an organisation that agrees on the technical standards and protocols needed to make the VO possible, acts as a framework for developing VO ideas and technology and for promoting them to the wider astronomical community and liaises with other disciplines facing similar data and metadata standardisation issues. These standards are there to *e.g.*, ensure uniformity in the data and metadata description for storage and accessibility; provide information about the production of datasets that can be used to assess their quality and reliability; enable tracing back the origin of datasets or documents (*e.g.* scientific articles, technical notes, etc.).

Data standardisation initiatives are certainly not unique to astronomy or natural sciences. Similar efforts are carried out in archaeology. The CIDOC Conceptual Reference Model (CRM; cidoc-crm.org), for instance, provides definitions and a formal structure for the description of concepts used in cultural heritage documentation and querying and exploitation of related dataset, with museums and museum collections as its main focus but with extensions towards field archaeology (Binding *et al.* 2008). Furthermore, data sharing, enabled by the development of standards and protocols is gaining momentum, as described below, with concepts such as interoperability infiltrating the field.

Data Sharing

Astronomy has been a pioneer in data sharing, with efforts like the database of the International Ultraviolet Explorer, a space-borne observatory that observed the Universe in the Ultraviolet (UV) part of the spectrum between 1978 and



4. ESASky (sky.esa.int) is an application that allows to visualise and download public astronomical data in science and explorer mode. View of the region around the nearby galaxy named NGC097. Access to images, catalogue and spectroscopic data are provided in visual form (*e.g.*, top left inset allows the selection of any catalogue data, listed in tabular form at the bottom of the panel). Clicking on any of the small squares (i.e., sources) on the image will trigger a popup with more information on the source. Data can be sent to other applications without local download, explored on-the-fly or exported on the local disk.

1996, emerging at the same time as the World Wide Web (Wamsteker *et al.* 1989). Nowadays, data sharing is a routine practice among astronomers and most datasets obtained by large astronomical facilities have a proprietary time of 12 months, after which they become publicly available via dedicated archives (Fig. 4). Astronomy, however, is only one among the many disciplines that are experiencing a strong push for data sharing and open access, by both their scientific communities and their funding agencies. Researchers in an increasing number of scientific fields are working together towards setting up interoperable data services and initiatives like the IVOA, such as the Research Data Alliance (RDA; Genova 2019; rd-alliance.org), are expanding across disciplines.

Archaeology has also been in the forefront of applying new and innovative techniques, still requests for data by e-mail are common practice, discipline-wide sharing is sporadic and only a small fraction of data is stored in structured formats that enable easy access (Marwick and Pilaar Birch 2018). Nevertheless, data sharing initiatives, often interweaved with standardisation works, are emerging with some notable examples including: EU-funded Advanced Research Infrastructure for Archaeological Dataset Networking (ARIADNE and its successor ARIADNEplus; ariadne-infrastructure.eu/portal/) that brings together existing archaeological data infrastructures to enable the incorporation of distributed datasets and new technologies into archaeological research methodologies, in a framework that has a lot in common with the VO; Open Context (opencontext.dainst.org), a system for sharing data in archaeology via the web for public use with a link to ORCID (orcid.org) for the identification of researchers;

the Digital Archaeological Record (tDAR; tdar.org), an international digital repository for long-term preservation; the Archaeology Data Service (ADS; archaeologydataservice.ac.uk), a digital repository for heritage data; or OpenDig (opendig.org), a site for direct online publishing of digital data for permanent storage hosting the Tall al-ʻUmayrī dig database, just to name a few.

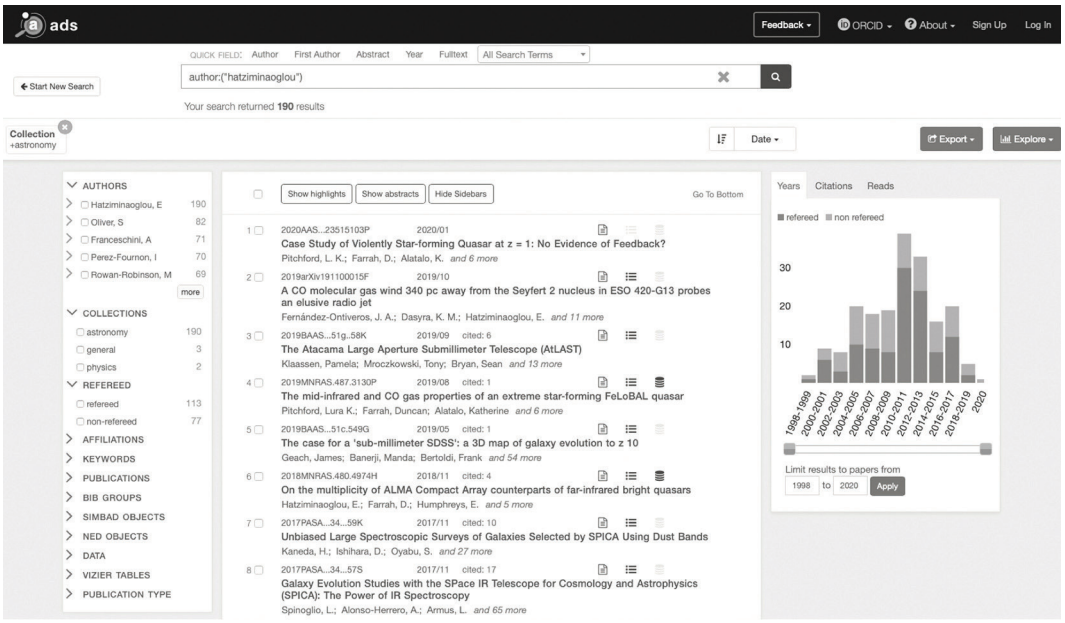
Data sharing is not a panacea for all data related issues. It is a practice that introduces concerns about quality assurance, metadata interpretation, data download tracking, provenance, referencing or acknowledgements, many of which are at the heart of the standardisation efforts mentioned above. Nevertheless, it comes with enormous benefits. Other than ensuring repeatability of the results and maximising the output of any scientific and research investments by enabling reusability of data for purposes beyond what was originally foreseen, data sharing also addresses many of the issues inherent to the handling of big data, regardless of discipline or scientific topic. It improves transparency and reduces duplication of efforts; it allows for new discoveries and facilitates science by less privileged countries/communities or scientist that do not have direct access to scientific facilities (*e.g.*, telescopes or archaeological sites in the vicinity). It also enables public participation and multidisciplinary work.

Open Access

Open access is gaining momentum under various models and configurations. The aim of this work is not to discuss such models nor the associated benefits or controversies, but rather to highlight ways preprints can be made publicly available, using once again the field of astronomy as an example.

If this were an astrophysics paper, I would have uploaded it, right after its acceptance and prior to its publication to a refereed journal, to arXiv (arxiv.org) (dubbed 'the archives'), an open-access platform for electronic prints. This effort started back in 1991 as an archive for preprints in physics but quickly expanded to astronomy, mathematics, computer science, quantitative biology, and, most recently, statistics, and it is currently maintained and operated by Cornell University. Submissions (made by registered users) are not peer-reviewed by arXiv itself, but the site is moderated to make sure that the submitted material have an academic content and are relevant to the category they have been submitted to. Preprints can be accessed by all internet users directly from the arXiv.org website or one from the many mirrors, without registration. Similar initiatives followed in other disciplines such as bioRxiv (biorxiv.org) for biological sciences operating since 2013 or PsyArXiv (psyarxiv.com) for psychological sciences operating since 2016.

A step further to such archiving efforts is provided by the Astrophysics Data System (ADS; ui.adsabs.harvard.edu/classic-form), a digital library portal for astronomy and physics, operated by the Smithsonian Astrophysical Observatory under a NASA grant. It provides access, via search forms, to bibliography databases (with a total content exceeding 13 million records) with publications in astronomy, astrophysics, and physics, including the arXiv. Other than direct links to the publications on the publishers' pages (access subject to fees), it provides free direct access to publications on arXiv, to abstracts, to bibliography statistics, citation counts, access to data linked from the publications (often via direct links to VO services), and other valuable resources (Fig. 5). The contents of the ADS are accessible to all internet users without registration. Finally, initiatives like ResearchGate (reseachgate.net), a social network primarily for sharing papers among researchers and scientists, provides open



5. Screenshot of the outcome of an ADS query for author "Hatziminaoglou" listing all publications of the author in reverse chronological order, providing a wealth of information on, *e.g.*, co-authors, journals, links to data, in the form of retractable menus (left), links to the abstracts and full papers (middle) and brief publication statistics (right), all clickable and interactive.

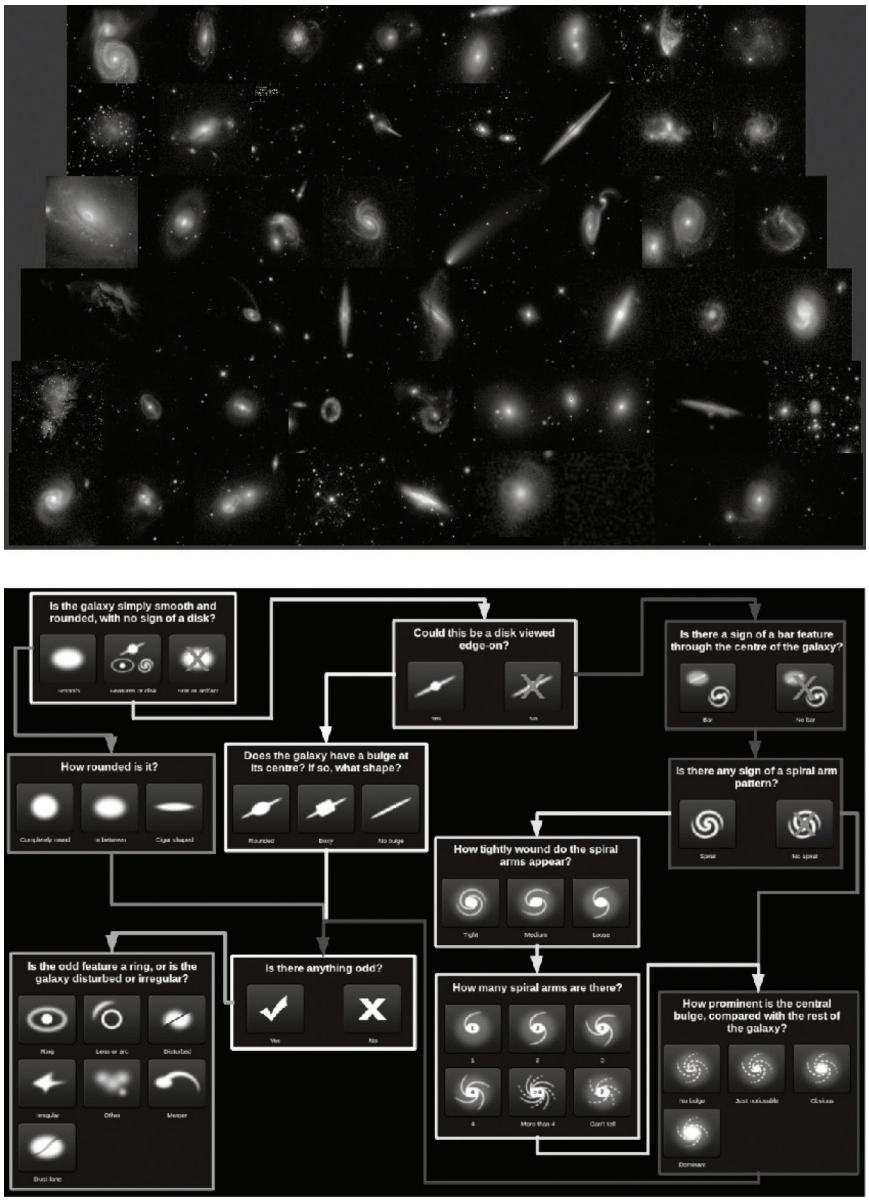
access to scientific papers from a wide variety of fields from one single platform, as well as opportunities for new collaborations via its multiple features.

People-Powered Research

One of the major breakthroughs of data sharing, and to an extent of open access, is that it opened the gates to people-powered research, allowing practically anyone with access to the internet to contribute to cutting-edge science and new discoveries.

Zooniverse (zooniverse.org) is the world's largest platform for people-powered research. It started off as a single project, Galaxy Zoo (initiated in 2007), whose task was the morphological classification of about 900,000 galaxies by eye, observed by the Sloan Digital Sky Survey (sdss.org) (SDSS; Fig. 6). More than 40,000,000 classifications were carried out by more than 100,000 volunteers in 175 days, that provided about 40 classification per galaxy in the 900,000 sample. Since then, Zooniverse has expanded to include more than 230 projects (about half of them currently active) in arts, biology, climate, history, language, literature, medicine, nature, physics, social science, and space. The creation of new projects can be done directly on the Zooniverse interface. The more than 1,000,000 volunteers from all over the world contributing to new discoveries that have been published in a stunning 238 peer-reviewed articles to date.

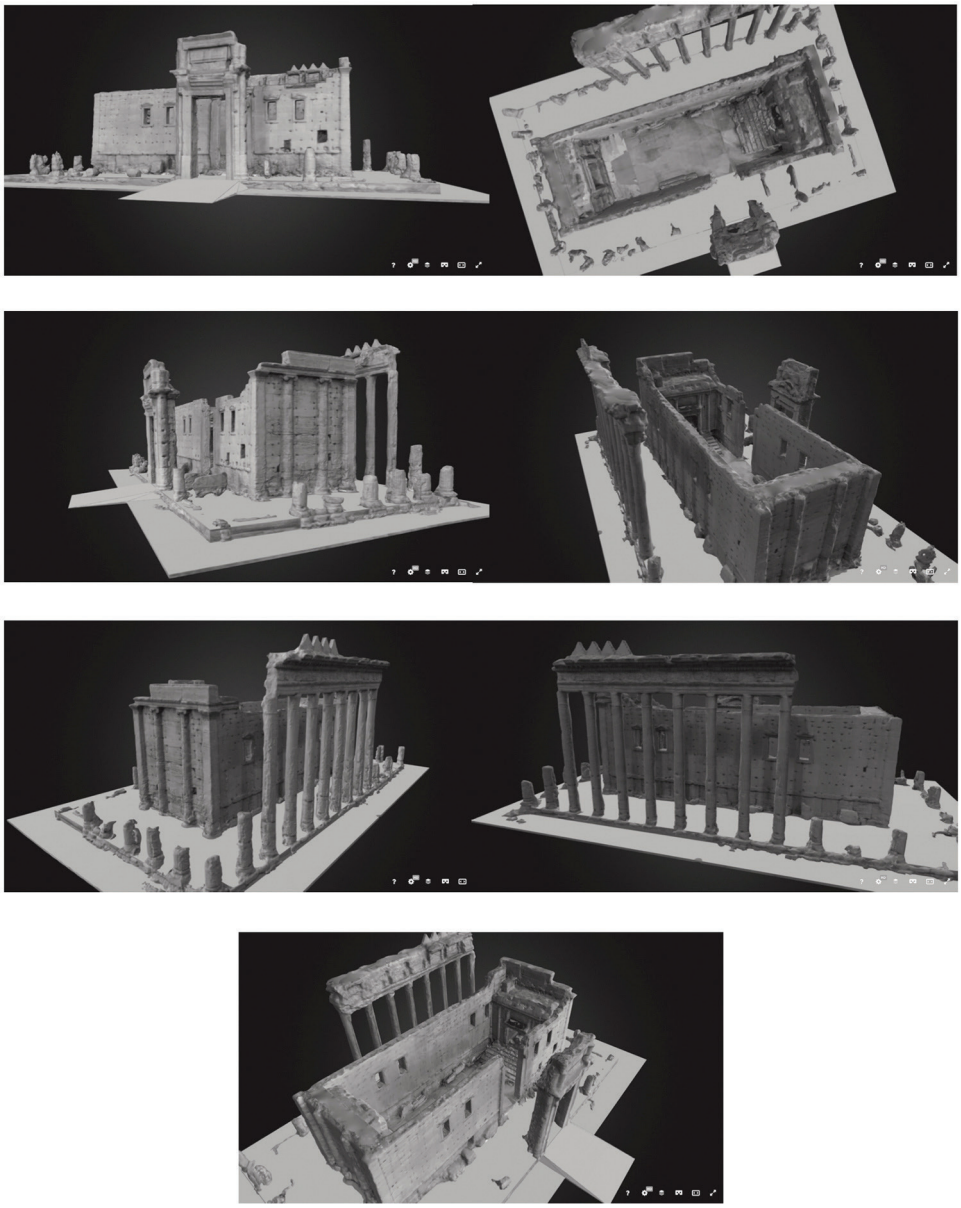
Earlier such initiatives include Clickworkers, a now defunct NASA project aiming at the identification of craters on Mars or Stardust@home that asks volunteers to look for the impact of interstellar dust particles on images of aerogel blocks exposed by the Stardust spacecraft after its launch in 1999.



6. Top panel: Variety in the morphology of galaxies observed by the Sloan Digital Sky Survey; bottom panel: Simple decision tree to be followed for the visual classification of each galaxy, by clicking on the interactive interface (Image credits: SDSS, Galaxy Zoo).

In astronomy, citizen science, a term whose definition largely depends on who you ask, comes in a variety of flavours other than crowdsourcing. Amateur astronomers, for instance, often act as comet or supernovae hunters, with many new discoveries made by their community every year.

Crowdsourcing efforts are becoming popular in archaeology, too. Currently paused GlobalXplorer (globalxplorer.org), is an example of a project that uses volunteers to analyse satellite images and look for



7. Different viewing angles of the model of Bel Temple in Palmyra (Syria), digitally reconstructed as part of Project Mosul.

new discoveries, but also signs of looting and encroachment in Peru and is based on a simple classification tree reminiscent of the one used by Galaxy Zoo. Furthermore, as datasets are often the only record of excavated or even destroyed sites, multiple efforts for digital reconstructions are taking place across the globe. Rekrei (projectmosul.org), for instance, formerly known as Project Mosul, launched in the aftermath of the destruction of the Mosul Museum by Daesh, is an

ingenious effort to digitally reconstruct lost heritage in the Middle East and elsewhere, by combining archaeology, photogrammetry, web development, and digital data (*i.e.*, photographs) taken by individuals (many of them tourists) over the years (Fig. 7). Similarly, the Zamani Project (zamaniproject.org) creates digital representations of historical sites in Africa that can be used for research, education, and restoration but also preservation for future generations.

Conclusions

The Square Kilometre Array (SKA; skatelescope.org) is the father of big data projects. The SKA is a global project that, towards the end of the decade, will be operating hundreds of dishes and several hundred thousands of low-frequency radio-antennas, all connected via the highest-speed network ever conceived for astronomical research. In early operations it is expected to produce an archive of science data products with a projected growth rate of a couple hundred petabytes per year (that is a couple times 1,000,000,000,000,000 bytes per year!). The needs of this project for data transfer, analysis, storage, and access demand continues change not only in the technologies used but also in the ways we think of and do science.

At the same time, new digital techniques are entering every step of the archaeological workflow causing an exponential increase in the data volume, opening endless possibilities for cross-disciplinary collaborations, perhaps even altering the profiles of professional archaeologists.

Although new efforts will build on the infrastructure already provided by earlier initiatives, they will require adaptation and technological leaps in order to accommodate the astronomical (pun intended!) data volumes involved. This is to say that the big data challenge is only starting. Data standardisation, sharing and interoperability efforts to bring together and curate data sets and other research resources across disciplines, to preserve their provenance, to render metadata searchable, and to make all those resources accessible as a whole, are taking up leading to a paradigm shift, allowing at the same time for new and unprecedented discoveries. This is the moment for disciplines to join forces towards developing common infrastructures and sharing knowledge, methodologies, and technologies.

This paper briefly discusses personal views of a professional astronomer with a profound interest in archaeology on how academic research is expanding not only beyond the traditional borders between individual disciplines but also beyond the strict limits of academia itself, how the combined power of the internet and the people around the world is employed resulting in new and exciting discoveries, how people with very different professional backgrounds and interests can contribute to original scientific research on topics distant from their fields of expertise, how technology can subscribe to the preservation and expansion of knowledge, and on how all of the above contribute to the globalisation of science.

Bibliography

Binding, C., K. May, and D. Tudhope. 2008. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC-CRM." In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, edited by B. Christensen-Dalsgaard, D. Castelli, B. Ammitzbøll Jurik, and J. Lippincott, 280–90. Berlin: Springer-Verlag.

Genova, F. 2019. "The Research Data Alliance: Building Bridges to Enable Scientific Data Sharing." *ASP Conference Series* 521:157

Hatziminaoglou, E. 2010. "Virtual Observa-

tory: Science Capabilities and Scientific Results." *ASP Conference Series* 424:411.

Marwick, B., and S.E. Pilaar Birch. 2018. "A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing." *Advances in Archaeological Practice* 6:125–43.

Szalay, A., and J. Gray. 2001. "The World-Wide Telescope." *Science* 293:2037.

Wamsteker, W. *et al.* 1989. "IUE-ULDA/USSP - The On-Line Low Resolution Spectral Data Archive of the International Ultraviolet Explorer." *Astronomy and Astrophysics Supplemental Series* 79:1–10.

Wells, D.C., E.W. Greisen, and R.H. Harten. 1981. "FITS: A Flexible Image Transport System." *Astronomy and Astrophysics Supplemental Series* 44:363.

Whitcher Kansa S., E.C. Kansa, and J.M. Schultz. 2007. "An Open Context for Near Eastern Archaeology." *NEA* 70:188–94.