Astronomical Images in the Light of Big Data

Ekaterina Postnikova[1], Natalia Chupina[1], Andrei Demidov[2], and Sergei Vereshchagin[1]

[1]Institute of Astronomy, Russian Academy of Sciences, Pyatnitskaya str., 48, 119017 Moscow, Russia
[2]Central Aerological Observatory, Pervomayskaya str., 3, Dolgoprudny, Moscow region, Russia
svvs@ya.ru chupina@inasan.ru es_p@list.ru the-admax@ya.ru

Abstract. The role of Big Data images in astronomy is considered, taking into account the usefulness of information in different periods of time. Estimates of the volume of accumulated data in the form of digitized photographic and CCD images are made. The rate of accumulation of information was analyzed. It is shown that only a small percentage of the obtained images at the present stage is already effectively used. We need the development of tools for intensive use and processing to obtain new knowledge. A convenient center for storing general information and simple search capabilities about the entire set of image archives, built on the principles of FAIR is important as part of the Virtual Observatory. The result of the initial development and creation of an image archive of the Zvenigorod Observatory is presented.

Keywords: Big Data, Images, Intensive processing of image archives

1 Introduction

Astronomy accumulated a huge amount of information in publications URL [1], whereas image archives are usually scattered at observatories and institutes. The images contain unique information about the Universe. This information is unrepeatable and can be used to solve scientific and other problems. The lack of a single center leads to the fact that even a professional is hard to find interesting and often necessary data.

With the help of astronomical photographic plates at the present time possible to open a new objects and phenomena. For examples, processing of photographic plates can be used to calculate the orbits of moving objects (asteroids, planets, satellites), as well as to study and search for variable stars, nova and supernova stars.

The aim of this work is determined from the problems of working with image archives carried out by us at the Zvenigorod Observatory. These are:

1) Improving the structure of the available archive of accumulated images obtained at the telescopes of the Institute of Astronomy RAS. It is important to testing the structure. The increasing volume of information and ability to combine photos and CCD images demand the development of using metadata. We need the change the scheme of his work in the framework of the principles of FAIR (Findable – Accessible – Interoperable – Reusable).

2) Estimate and use of the current world practice of storing observation archives in the field of images of objects of the starry sky. Also, develop and use of options for the

specific storage of archives of various volumes both at small observatories and at the largest observatories.

3) Development of a way to include archives into service of the open access and data retrieval in the totality of astronomical information that is part of the World data.

2 Photo Images

For more than 100 years, being the pre-CCD era, a large amount of images obtained by the photo, what was the main method for that period have accumulated. Although many observatories did not carried out large-scale observation in the modern sense, the accumulated material yet has large volume, including by the large number of observatories and telescopes. We will begin consideration with the well-known observatory of our institute and proceed to the archives of other observatories and, then, to modern electronic data. More than 100 astronomical observatories operating around the world have similar archives of images taken on photographic plates. With the help of photographic plates images have discovered the satellite of Pluto - Charon Christy and Harrington [2]. Similarly, now the well-known planets outside the Solar System (exoplanets) were first found in the photo image by Farihi and Stoop [3].

3 From Photo Observations to Electronic Images at the Zvenigorod Observatory

Photographic images are used along with digital ones. For this it is necessary to convert them to digital format. For deriving astrometric solutions, object identification and photometry from digitized photo plates the systems like URL [4] – astrometry.net are used, other one – SExtractor [5]. They include photo-observation, conversion photo into electronic form and observation with CCD.

We have 4500 photo plates. Information from each plate is allocated in the amount of 1 GB. In total, it turns 0.0045PB. The observatory has two large telescopes. ASTROGRAPH ZEISS-400 with lens diameter D=40 cm and VAU CAMERA with mirror diameter D=107 cm. More information and details can be found in URL [6].

Photographic survey of the Sky (FON [7]) and FOCAT [8] programs) was carried out from the 1980 to 1992 ys. The FON and FOCAT programs also were used to determine the coordinates of 113 Galactic radio sources [9].

The most interesting images are in our archive of comets. Its distinguishing features are the long series of observations of the brightest comets of the late 20th century (Halley, Hale-Bopp, Hyakutake, etc.). The Zvenigorod archive has 220 images for 11 comets URL [10]. In Fig. 1, we see an example of an image of Hyakutake comet.

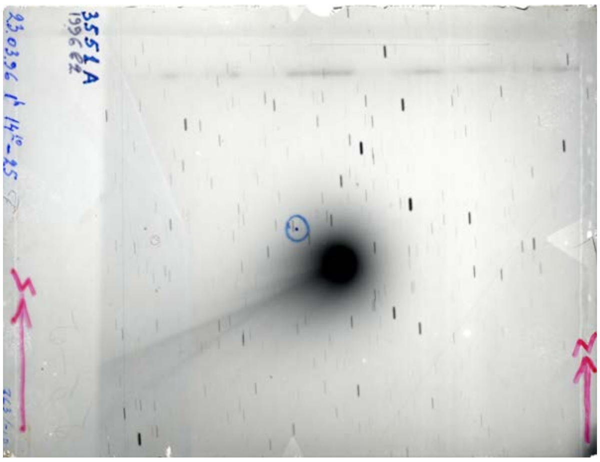

Fig. 1. Comet Hyakutake. Observation at Zvenigorod Observatory. Date of observation: 03/23/1996. Plate Number: 3551A. Plate size: 13 × 18 cm. Obtained by V.P. Osipenko

We have 4500 photo plates, or 4.5 TB. Zvenigorod Observatory can be considered to be similar in terms of information to about a hundred more observatories worldwide. This value will increase two orders and will be approximately 0.5 PB. This, it is easy to understand, is approximately the total amount of photographic images accumulated on the small observatories of the World.

The results of modern observation with Zvenigorod robot-telescope are in digital format, Fig. 2. The stream from the robot-telescope of the one image takes 32 MB. Photometric observations are carried out, on average, with an exposure of 60 – 120 seconds. From 30 to 60 images are recorded in one hour. Given that the night lasts an average of 8 hours, we get 23 GB for one night. In our area in the year 27% of nights are observant, that is, about a hundred nights per year. During this time, you can gain 2.3 TB of information.

Robotic wide-angle (field of view is 10°) system is used for the sky monitoring at the Zvenigorod Observatory INASAN. It makes possible the automatic observations according to a predetermined plan. The telescope robot is based on a wide angle telescope and photo detector with U, B, V, R, I photometric system, Terebizh [11]. The angular diameter of the field of view is 10° for the matrix 50 × 50 mm, the focal length is 395 mm. It has direct drive, [12]. Extending the knowledge of the Zvenigorod Observatory to other ones, we can say that total number of CCD images at all small observatories of the World will be approximately 0.2 PB.

Fig. 2. The LO Peg is a young star of the K3 spectral class and is one of the most studied fast-rotating stars of the late spectral classes. The star is a member of the AB Dor [13] group of stars with common spatial motion. The star gave its name to the moving group AB of Golden Fish, a stellar association consisting of approximately 30 stars that move in the same direction and are approximately the same age. The image was obtained by S.A. Naroenkov and M.A. Nalivkin

4 Large Surveys, Observatories and Telescopes

The **Pan-STARRS** (Panoramic Survey Telescope and Rapid Response System) URL [14] automatic system consist of four Richie-Chretien telescopes with mirrors of 1.8 meters each and 1.4-gigapixel CCD cameras located on the top of Mauna Kea volcano on the island of Hawaii. Its archive size is 1.6 PB, which makes it the largest astronomical data base ever released.

Gaia data. A large space project currently underway is a Global Astrometric Interferometer for Astrophysics (Gaia) URL [15]. Now the second Gaia data, Gaia DR2, have released for approximately 1.7 billion sources brighter than magnitude 21. First Gaia realize (DR1) consist 1 billion row or 351GB of data and the second Gaia DR2 consist 1.7 billion rows or 1.2TB of data.

BTA-6. The Special Astrophysical Observatory is located at a height of 2070 meters in Karachay-Cherkessia. Mirror-reflector 605 cm, azimuthal mount, [16]. The SAO has a “General Observational Data Archive” URL [17]. Roughly it exceed 1 TB. The Harvard College Observatory Astronomical Plate Stacks URL [18]. There are over 500,000 glass photographic plates (memory volume will be a 500 GB).

Very Large Telescope (VLT) is located in the Atacama Desert, mountain Paranal, 2635 m (Chilean Andes). And it belongs to the European Southern Observatory (ESO), which includes 9 European countries. A system consists of four telescopes 8.2 meters and four auxiliary 1.8 meters. According to luminosity, is equivalent to one device with a mirror diameter of 16.4 meters This telescope generates an average of 20–25 gigabytes of information per night URL [19]. The all archive value equal no more than 25 TB.

Large Binocular Telescope (LBT) is located on Mount Graham, at an altitude of 3321m, in Safford, Arizona (USA). Two mirrors of 8.4 meters with a distance of 14.4 meters, which is in terms of sensitivity is equivalent to one mirror with a diameter of 11.8 m. According to estimates, two telescopes of 8.4 meters both generate 92 GB / night of data URL [20]. The all archive value equal no more than 90 TB.

The Gran Telescopio CANARIAS (GTC) stands on top of the extinct volcano Muchachos on one of the Canary Islands, at an altitude of 2396 m, [21]. The diameter of the main mirror is 10.4 m. It generate up to 20 MB of data per second. In the course of a typical night, therefore, it is possible to accumulate up to 720 GB.

Hobby-Eberle telescope (HET) has segmented 11x10 m mirror (effective area 9.2 meters), located in the USA (Davis Mountains – 2026 m, Texas). It is equipped with active optics [11]. The data volume was obtained on this instrument at 120 GB / night and 20 TB in a three year survey [22].

Hubble Space Telescope (HST) was launched in 1990 and should work until 2030. The diameter of its mirror is only 2.4 m, URL [23]. For 29 years, HST has generated 153 Tb of data.

As a result, a separate position is occupied by the PAN-STAR archive with a volume of 1.6 PB. The rest of the total does not exceed 0.35 PB.

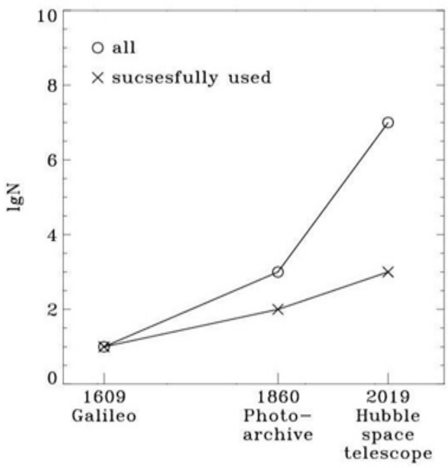

Fig. 3. The increase of information over time. From the Galileo telescope to the Hubble Space Telescope. Small circle shows the logarithm of a number of images (N) representing a full amount of data, and crosses – the volume of data that gave new results. We see that the effectiveness of the “archives” of images is now far from the same as it was before. If Galileo made a discovery from almost every image, then a modern telescope gives only one useful image out of 100,000. The data presented in this figure are estimates only

5 The Increase of Requests

On Fig. 3 the growth of information with time and the change in the efficiency of its use are shown. How many discoveries does astronomical PB contain? Is it true to say that the percentage of images giving new knowledge fell from 100% Galileo to 10 percent in modern observations? If in the search engine of the Digital Library for Physics and Astronomy ADS URL [1] we make a request to get the number of publications with the words “image” and “image&archive”, then we get a number, the dependence of which on the dates is shown in Fig. 4.

As we can see, in Fig. 4 there are no publications about the archives of images before 1980. This is due to the history of digital technology. The CCD was invented in 1969 by Willard Boyle and George Smith at Bell Laboratories (AT & T Bell Labs) URL [24]. In 1970, Bell Labs researchers learned how to shoot images using simple linear devices. Soon these devices appeared as light receivers on telescopes. And since the end of 80-s, publications about archives began to appear.

Just about the images of stars, it was almost always, starting with the book of Ptolemy [25]. Although the rapid increase in the number of publications began precisely on photographic images.

In 1839, on January 7, physicist Francois Arago, at a meeting of the Paris Academy of Sciences, first reported on the invention of daguerotype [26] by Louis Daguerre and Nicephorus Niepce. By the decision of the IX International Congress of Scientific and Applied Photography, this date is considered the day of the invention of photography. It is after this that we see a rapid increase in the number of publications on the processing of astronomical photographic images in Fig. 4. As we have said, the number of publications on archives has dramatically increased since the late 80s. Looking at Fig. 4, we will see that the graph by the number of publications with the words “image” in the headline also greatly increased its inclination from the same time.

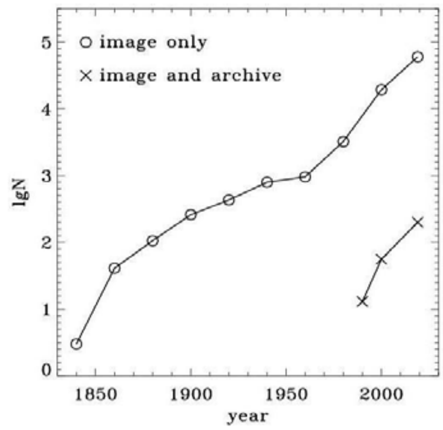

Fig. 4. The cumulative distribution of the number of publications with “image” in the title (small circles) and “image&archive” (slanting crosses)

We propose to consider the efficiency of the telescope (the coefficient of efficiency QE). Obviously, it is determined by the diameter of the mirror / lens -D (the image quality depends primarily on D) and the speed of information accumulation (delta t, time spent on accumulation). We get QE=D VTB / delta t. In Fig. 5 shows the dependence of QE on D.

Table 1. The QE values for different telescope

Telescope	D, m	VTB, in TB	Δt, yr	QE /(TB/yr)	lg QE
ZEISS-400, Zvenigorod	0.40	4.5	30	0.060	–1.20
Zvenigorod robot-telescope	0.25	2.3	1	0.575	–0.24
Gaia	4.80 eff	1.2	5	1.150	0.06
Hobby-Eberle telescope (HET)	10.00	20	3	66.7	1.82
Greater Canary Telescope, GTC	10.40	240	1	2496	3.40
Hubble	2.40	153	28	13.1	1.12

The effectiveness of small telescopes, as we see in Fig. 5, differs up to a thousand times. But even among large telescopes, the QE scatter is large. Table 1 shows data for various telescopes: mirror on lens diameter, volume of accumulated information (VTB), the time interval of data accumulation, QE value and log QE. In order to try to estimate the amount of information extracted from the universe by any telescope, we came up with an “image production coefficient,” or QE. Let us make the following assessment for ZEISS-400. About thirty years of operation were received 4.5 TB information on 40 cm lens. We get a one year by one centimeter of telescope mirror (objective lens for a refractor, as in this case) image output in bytes QE, or QE for Astrograph will equal 0.06 TB/(yr cm).

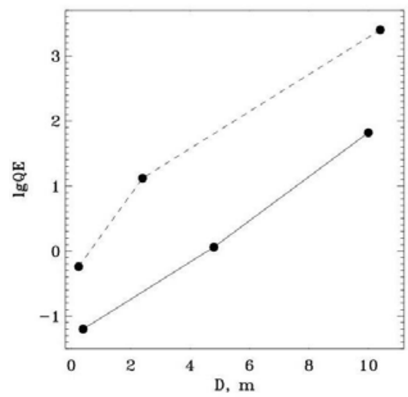

Fig. 5. The logarithm of the efficiency coefficient of the telescope depending on the diameter of the mirror (or lens)

It would seem that the accumulation time is an important parameter to get the value efficiency, but, as determined empirically, it is mainly determined by the diameter of the telescope, Fig. 5.

6 Importance of Format, Problem of Identifying

The problem of identifying stars and other objects cannot always be successfully solved by using automatic identification programs [27]. In our archive the most interesting images obtained by long series of observations of comets, Fig. 6. In such cases the images of stars are obtained extended, which greatly complicates their identification.

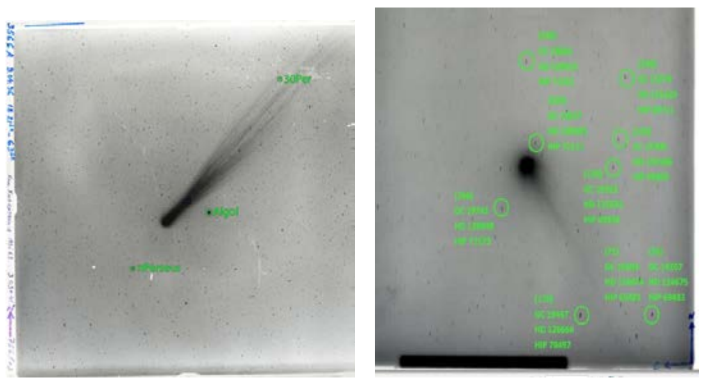

Fig. 6. Examples of the identification of objects on photo plates. On the left we see an extended object – Hyakutake comet. As the telescope was guided on the comet, the images of the stars were stretched. Stretching images of stars will create difficulties for automatic identification. Identification by several catalogs is shown on the right plate – it is Hipparcos, HD, GC, and Star Atlas. The plates were obtained by V.P. Osipenko, identified by V.P. Osipenko and M.D. Sizova

7 Inclusion of Astronomical Image Archives in the World Data Center

There are principles for working with data – FAIR Data Principles URL [28]. These data principles (stands for findability, accessibility, interoperability, and reusability) are a set of guidelines to make data searchable, compatible and reusable [29].

It is necessary to create reliable repositories of data, related audit and certification schemes (for example, Core Trust Seal URL [30] gives both repository requirements and a list of the most reliable ones). These demands are from resulting from the accession to the Data Certificate (DSA) and the certification scheme under the auspices of the Research Data Alliance (RDA, established to: ensure data sharing, overcoming technological, national and disciplinary barriers).

8 Conclusions

The total memory capacity of computer-based storage media required for astronomical archives – 1.6 PB from Pan-STARRS and 0.35 PB for other large telescopes and observatories. The share of small observatories accounts for approximately 0.5 PB photo archives and 0.2 PB for the CCD. In sum, how easy it is to see, 2.65 PB.

It is important to consider the possibilities of connecting to the World Data System. While it is not traditional for astronomical data. WDS URL [31] is building worldwide ‘communities of excellence’ for scientific data services by certifying Member Organizations – holders and providers of data or data products – from wide-ranging fields by using internationally recognized standards. WDS Members are the building blocks of a searchable common infrastructure, from which a data system that is both interoperable and distributed can be formed. This path, if it is actively supported financially by the government, puts the work with astronomical data before a real choice towards overall improvement.

Results. 1) The result of the development and creation of an image archive of the Zvenigorod Observatory is presented.

2) Astronomical archives of various levels from the small observatories to the largest telescopes were studied. It is shown that today astronomical images are in archives by different structure and access, both at many small observatories and at the largest specialized network resources. A significant part of the images is included in publications, which causes copyright problems. The possibilities of solving the copyright problem developed by the scientific community are shown.

3) A convenient center for storing general information and simple search capabilities about the entire set of image archives, built on the principles of FAIR, is important as part of the Virtual Observatory.

References

1. Astrophysics Data System (ADS), http://adsabs.harvard.edu/abstract_service.html.
2. Christy, J. and Harrington, R.S.: The satellite of Pluto. Astronomical Journal **83**, 1005, 1007, 1008 (1978). http://articles.adsabs.harvard.edu/pdf/1978AJ.....83.1005C
3. Farihi, J. and Stoop, J.: Extrasolar planetary systems were first observed a century ago, evidence suggests Are we celebrating the 20th or 100th anniversary of exoplanet discoveries? https://www.elsevier.com/connect/extrasolar-planetary-systems-were-first-observed-a-century-ago-evidence-suggests
4. Astrometry.net software, http://astrometry.net.
5. Bertin, E. and Arnouts, S.: SExtractor: Software for source extraction. Astronomy and Astrophysics Supplement **117**, 393–404 (1996).
6. Equipment of Zvenigorodskaya observatory, http://www.inasan.ru/en/divisions/zvenigorod/instr/
7. Kolchinsky, I.G. and Onegina, A.B.: On the Programme of Sky Photographing with Wide-Angle Astrographs Astrometriia i Astrofizika **39**, 57 (1979).
8. Bystrov, N.F., Polojentsev, D.D., Potter, H.I., Yagudin, L.I., Zallez, R.F., and Zelaya, J.A. Bulletin d'Information du Centre de Donnees Stellaires **44**, 3 (1994).
9. Rizvanov, N., Dautov, I., and Shaimukhametov, R.: The comparative accuracy of photographic observations of radio stars observed at the Engelhardt Astronomical Observatory. A&A 375, 670–672 (2001).
10. Archive of photoplates scan of the Zvenigorod Observatory INASAN with images of comet Hyakutakehttp://www.inasan.ru/divisions/zvenigorod/scan/scan_hyakutake_comet.
11. Terebizh, V.Y.: On the Capabilities of Survey Telescopes of Moderate Size, Astron. J. 152, 121 (2016).
12. Savanov, I.S., Naroenkov, S.A., Nalivkin, M.A., Puzin, V.B., and Dmitrienko, E.S.: Photometric Observations of LO Peg in 2017, Astrophysical Bulletin **73** (3), 344–350 (2018).
13. Zuckerman, B. and Inseok Song: The AB Dorados moving group, The Astrophysical Journal, (613), L65–L68 (2004).
14. The Pan-STARRS1 data archive home page, https://panstarrs.stsci.edu/
15. Gaia Archive, http://gea.esac.esa.int/archive/
16. BTA-6 telescope hamepage, http://w0.sao.ru/
17. General Observational Data Archive, https://www.sao.ru/oasis/cgi-bin/fetch?lang=ru
18. The Harvard College Observatory Astronomical Plate Stacks, http://tdc-www.harvard.edu/plates/
19. Very Large Telescope (VLT) homepage, http://www.eso.org/sci/facilities/paranal/telescopes/vlti.html
20. Large Binocular Telescope (LBT) homepage, http://oldweb.lbto.org/
21. The Gran Telescopio CANARIAS (GTC) homepage, http://www.gtc.iac.es/
22. Hobby-Eberle telescope (HET) homepage, http://www.as.utexas.edu/mcdonald/het/het_gen_01.html
23. Hubble Space Telescope (HST) homepage, http://hubble.nasa.gov
24. Boyle, W.S. and Smith, G.E.: Charge Coupled Semiconductor Devices. Bell Syst. Tech. J. 49 (4), 587–593 (1970).
25. Ptolemaeus, Claudius Astronomia, teutsch Astronomei: von art, eygenschafften, und himelischen Bildern und iren Sternen wirckung der XII Zeychen des Himels,der VII Planeten, der XXXVI himelischen Bildern und iren Sternen, by Ptolemaeus, Claudius, 1545. (1545). DOI: 10.3931/e-rara-1983.

26. Arago, François: Le Daguerréotype. In: Comptes rendus IX (July–Dec. 1839): 250-67. 4to, 903. Paris: Bachelier, 1839.
27. Automatic stars identification on astronomical images, http://nova.astronet.com
28. FAIR Data Principles, https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf
29. Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18.
30. Core Trust Seal, https://www.coretrustseal.org/.
31. The World Data System (WDS), Interdisciplinary Body of the International Science Council (ISC; formerly ICSU), URL: https://www.icsu-wds.org/organization