A multi-dimensional dataset of Ordovician to Silurian graptolite specimens for virtual examination, global correlation and shale gas exploration

Hong-He Xu [1*], Zhi-Bin Niu [1,2*], Yan-Sen Chen [1], Xuan Ma [1], Xiao-Jing Tong [1], Yi-Tong Sun [1], Xiao-Yan Dong [1], Dan-Ni Fan [1], Shuang-Shuang Song [1], Yan-Yan Zhu [1], Ning Yang [1], Qing Xia [1]

[1] State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology and Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, 210008 Nanjing, China

[2] College of Intelligence and Computing, Tianjin University, 300354 Tianjin, China

*The first authors.

Correspondence: Hong-He Xu (hhxu@nigpas.ac.cn), or Zhi-Bin Niu (zniu@tju.edu.cn)

Abstract

Multi- elemental and -dimensional data are more and more important in the development of data-driven research, as is the case in modern palaeontology, in which examination~~s,~~ by experts, or someday ~~the~~ artificial intelligence, ~~to~~ every fossil specimen plays~~acts~~ a fundamental role. We here release a~~n~~ ~~integrated~~ dataset of 1,550 graptolite specimens representing 113 Ordovician to Silurian graptolite species or subspecies that are significant in global stratigraphic correlation and shale gas exploration. The dataset contains 2,951 high-resolution images and a ~~structured~~ data table of each specimen's scientific information, e.g., ~~every specimen's~~ taxonomic, geologic, and geographic information, ~~comment,~~ and references. ~~Specimen data of~~ ~~o~~Our dataset provides images ~~virtual examinations~~ for specialists or laymen worldwide, is supported ~~are visualized,~~ by the tool ~~we developed~~, FSIDvis (Fossil Specimen Image Dataset Visualizer), which we developed to facilitate the human-interactive exploration of the rich-attribution image dataset, and also ~~are analysed with~~ a nonlinear dimension reduction technique, t-SNE (t-Distributed Stochastic Neighbor Embedding), to project image data into ~~the~~ two-dimensional space to visualize and explore ~~the~~ similarities. Our dataset potentially contributes to virtual examinations of~~to~~ specimens (VES), global

bio-stratigraphic correlation, and improvement of the shale gas exploration efficiency. ~~A fossil specimen database need to fulfil the purpose and the requirement of VES.~~ All data, images and the spreadsheet file, are available from https://doi.org/10.5281/zenodo.5205215 (Xu, 2022).

1. Introduction

Fossils are ~~show the~~ direct evidence of prehistoric life and are probably the most important research object of palaeontology and stratigraphy, during which fossils are collected, sampled, illustrated, described, curated, and deposited as permanent specimens in museum or institution for further investigation (Shute and Foster, 1999). Examinations to fossil specimens is a key and indispensable part of ~~in~~ descriptive~~onal study of~~ palaeontology. Such, however, can be partially achieved in a convenient and low-cost way, with the aid of multi-dimensional fossil specimen dataset as in this study.

Graptolites are ~~is~~ an extinct group of marine, colonial, organic-walled hemichordates and have~~s~~ over 210 genera/3,000 species in worldwide fossil records from the Cambrian to Carboniferous (c. 510~320 Ma) shales ~~sediments~~ (Maletz, 2017). Graptolites extensively diversified in the Ordovician Period and witnessed the second-largest mass extinction in geological life history, i.e., the end-Ordovician mass extinction (Goldman et al., 2020). Graptolites evolved quickly and spread globally in the Paleozoic (Fig. 1), and its species are widely used as significant index fossils for determining rock ages and regional bio-stratigraphic correlation. Bio-zones based on graptolite species dividing the Ordovician and Silurian Periods~~sediments~~ are generally less than one million years in duration; such a short geological interval~~moment~~ makes ~~it~~ possible ~~for~~ a precise understanding of life evolution in geological history (Chen et al., 2012; 2018). Up to 102 Ordovician and Silurian graptolite species were selected as global bio-zones for dating sediments and understanding the evolutionary pattern of palaeobiology; and 13 global stratotype sections and points (GSSPs) are defined by the first appearance datum (FAD) of graptolite species from the Cambrian, Ordovician, and Silurian systems (Goldman et al., 2020). (Fig. 2).

Additionally, bio-zones or indication zones based on graptolite species assist with identifying mining beds for shale gas exploration (Fig. 1). Graptolitic~~e~~ shale yields a significant volume of shale gas and comprises more than 9% global hydrocarbons rocks (Klemme and Ulmishek, 1991; Podhalańska, 2013). In China, over 61.4% of natural gas is yielded from~~ the~~ Ordovician and Silurian graptolitic~~e~~ shale of southern China (Zou et al., 2019). Identification of graptolite species helps to locate shale gas mining beds; especially, 16 graptolite species were chosen as “gold callipers” to locate favourable exploration beds (FEBs) of shale gas from China (Zou et al., 2015) (Fig. 2).

In this paper, we describe a multi-dimensional and integrated dataset of graptolite specimens. The dataset potentially contributes to a range of scientific activities and provides 1) ~~an ~~easy access to and ~~the ~~virtual examination of~~to~~ fossil specimens through high-resolution images and detailed scientific information for teaching and training in paleontology and geologic survey; 2) a standard fossil specimen image dataset for use~~d~~ in bio-stratigraphic correlation and to improve exploration efficiency in the shale gas industry, and 3) a potential aid of developing image-based automated classification~~ model~~.

2. Materials and methods

All images in~~of~~ our dataset were taken from graptolite specimens that are preserved in~~as~~ shale and were collected from China. These specimens are housed at the Nanjing Institute of Geology and Palaeontology (NIGP), Chinese Academy of Sciences (CAS), with serial numbers and prefix NIGP.

We spent over two years to photograph every specimen using a single-lens reflex camera Nikon D800E with Nikkor 60 mm macro-lens and a Leica M125 or~~and~~ M205C microscope~~s~~ equipped with Leica cameras (Fig. 3). Every image is well focused and ~~better ~~shows the morphology of the graptolite~~-bodies~~. In total, we took 40,597 images, including 20,644 camera photos (each with a resolution of 4,912 × 7,360) and 19,953 microscope photos (each with a resolution of 2,720 × 2,048). Photos of low contrast or bad focus were removed from the whole collection. We ~~only kept and ~~selected only~~the~~ photos that show the morphology of the ~~every ~~specimen and the diagnostic characters of each graptolite species that the specimen represents (Fig. 4). We selected one or two images for each specimen as the ~~present ~~final dataset, uploaded to, and stored in our cloud server (Fig. 3). ~~Every specimen has at least one original photo, and another image shows specimen with a~~

~~scale bar. Occasionally in some cases of large image, the scale bar is embedded, just beside the fossil itself.~~

3. Data description

Our final dataset consists of 2,951 high-resolution images and a related spreadsheet file. Every image is a high-resolution photo taken from a collection of 1,550 graptolite specimens. These specimens were formally published between~~in~~ 1958 and~~-~~ 2020. They~~, and taxonomically~~ belong~~ing~~ to 113 graptolite species or subspecies~~,~~ of 41 genera and 16 families of the Order Graptoloidea (see the spreadsheet file, Fig 5). The geological age of these graptolite species ranges from the Middle Ordovician ~~to~~ (467.3 Ma) to the Telychian (433.4 Ma) Stage of the Silurian Period (Fig. 5).

These graptolite species have relatively abundant fossil records and are significant in regional and global bio-stratigraphic correlation. They are commonly used in geological age determination and shale gas FEB indication, including 32 graptolite bio-zones from the Darriwilian Stage of the Ordovician Period (467.3 Ma) to the Telychian Stage of the Silurian Period (433.4 Ma) and 16 “gold callipers” of shale gas FEBs for the case of 20 m to 80 m thick graptolite shale in China (Fig. 6). These species also include two “golden spike” graptolite species for the two GSSPs in southern China (i.e., bases of the Darriwilian Stage in the Middle Ordovician System and the Hirnantian Stage in the Upper Ordovician System)(Goldman et al., 2020; Zhang et al., 2020).

The name of the individual image file is initialled by the specimen’s unique number and taxonomical species name. Every specimen was photographed with scale bar. The scale is attached to an image of the entire rock specimen. The other image is a close-up of the fossil within the coloured loop drawn on the whole specimen. Occasionally in the large images, the scale bar is embedded and beside the fossil specimen. For example, in the file named ‘9721Cardiograptus_amplus_S.jpg’, the genus name and species name are connected by the underline symbol, avoiding the space symbol. ‘9721’ is the specimen number, ‘Cardiograptus_amplus’ means the species name is *Cardiograptus amplus* and ‘_S’ means it is a photo with scale bar. In all scale bars, the minimum unit is one millimetre.

The image files are ~~is~~ in JPG format. The single JPG file size ranges from

822 KB to 7.055 MB. The whole volume of the dataset is 10.4 GB. The quality of specimen images in our dataset is much better than that in any previous publications because ~~version for that~~ most specimens were first~~ly~~ studied many years ago and their illustrations were in black and white, in low-resolution and/or printed on paper publications only. Most of these specimens were illustrated only once, or never clearly photographed. The image collection of our dataset provides necessary complement for these specimens and, furthermore, ~~once again~~ unfolds their scientific value to experts or anyone who is interested in ~~with~~ fossils.

Every ~~piece~~ of specimen is tagged with scientific information, including genus and species names, nominator, nomination year, specimen number, collection number, locality (province, city, county), geological horizon and section, collector name, collecting time, identifier, identifying time, related references, and ~~published~~ illustration labels. Specimens can be indexed and located in their detailed housing drawers and cabinets using any of the above information. Their detailed geologic ~~research-related~~ information can also be obtained from the geological section-based database, the Geobiodiversity Database (Xu et al., 2020) and forms key elements of fossil specimen metadata (Xu et al., in press). All related information is collected and recorded in a separate spreadsheet file released with our image dataset (Xu et al., 2022).

~~Additionally, considering s~~Some specimens of our collection have a long research history, since 1958, and their taxonomical status might have changed in the ~~new~~ light of graptolite systematic studies~~y~~ (Maletz, 2017; Zhang et al., 2020).~~,~~ W~~w~~e invited graptolite palaeontologists to curate every specimen to make sure that its scientific information is updated and widely accepted. ~~The comments, as emendation results, are also showed in the spreadsheet file of our dataset.~~ The spreadsheet file includes following fields: species ID, Phylum, Class, Order, Suborder, Infraorder, Family, Subfamily, Genus, Revised species name, tagged species name, total number of specimens, specimen serial number, image file name, microscope photo number, SLR photo number, Stage, a~~A~~ge from, a~~A~~ge to, mean age value, locality, longitude, latitude, horizon, and ~~specimen~~ first~~ly~~ published reference. It is noted that the ‘Revised species name’ of every specimen reflect the emendation and correction study in Ma (2020~~1~~), with help of graptolite expert

Prof Zhang Y-D (NIGP), which might need further study or peer-reviewed. One can always search specimens according to their tagged species names.

Our dataset, with the image collection and comprehensive information of a large batch of fossil specimens, provides supports virtual examinations of to specimens in a convenient and low-cost way. Experts or laymen can look through, examine, study, and even measure fossil specimens without need for regional/international travel and formalities. Such greatly benefits palaeontology in research, teaching, and science communication (Rahman et al., 2012).

4. Data visualization

We have developed an interactive web exploration tool, FSIDvis (Fossil Specimen Image Dataset Visualizer), to assist users to examine better the scientific contents of our data (Fig. 7).

We further explore the distribution of these graptolite images and visualize the t-SNE feature embedding of our graptolite dataset (Fig. 8) using different colors to denote different specimensfamilies. In detail, for each annotated image, we first resized it into 448×448 pixels and fed it into the trained Convolutional Neural Network (CNN) model. The output 1×1×2048 feature map from the last average pooling layer is flattened and projected to a 113 (number of species) dimensional fully connected layer to represent an image embedding. After that, we use t-SNE (t-Distributed Stochastic Neighbor Embedding), a nonlinear dimension reduction technique for high-dimensional data, to project the image embeddings into the two-dimensional space for visualization. Finally, we indicate the image data distribution by a scatter plot, we use 15 colors to represent 15 families of the order Graptoloidea, covering 42 genera and 113 species., so Tthe distribution of the images in this figure is based on species, which showing s a potential of automatic classifying graptolite species using artificial intelligence (Niu and Xu, 2022) "big mixed, small settlements" posture.

5. Conclusions

A multi-dimensional, integrated dataset based on 1,550 pieces of graptolite specimens is released. It contains 2,951 high-resolution images and a spreadsheet file showing structured records of every specimen's scientific information. During the preparation of the dataset, 113 Ordovician to Silurian graptolite species or subspecies were selected for their significances in stratigraphic correlation and shale gas exploration, and these specimens were carefully photographed and taxonomically curated.

Our dataset provides experts or laymen with a mean of virtual examination of ~~to~~ a batch of fossil specimens in a convenient and low-cost way. It potentially contributes to global bio-stratigraphic correlation, especially with those bio-zone graptolite species, and in the shale gas industry to improvement of exploration efficiency. A fossil specimen database needs to fulfil the purpose and ~~the~~ requirement of virtual examination of~~to~~ specimens.~~,~~ This~~such~~ greatly benefits palaeontologic~~y~~ research and science communication.

The whole dataset is visualized by the tool FSIDvis (Fossil Specimen Image Data Visualizer) and a nonlinear dimension reduction technique, t-SNE (t-Distributed Stochastic Neighbor Embedding)~~, showing their potential using in automatic classifying in the future~~.

Data availability. The dataset is archived and publicly available from https://doi.org/10.5281/zenodo.5205215. Visualized version is available at http://fsidvis.fossil-ontology.com:8089/

Author contributions. H.-H.X. and Z.-B.N. equally designed the project, developed the model, and performed the simulations. H.-H.X. prepared and revised the manuscript. Y.-S.C. gave technical~~ian~~ support~~s~~. X.M. ~~revised and~~ curated fossil specimens. Others contributed to ~~in~~ specimen photography.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgments. We thank Prof. Zhang Yuandong and Dr. Chen Qing, NIGP, CAS, for careful curating and examining of the ~~to~~ graptolite specimens; Prof. Peter M Sadler, University of California (USA), for comment and improving the manuscript; Dr. Pan Zhaohui, Institute of Vertebrate Palaeontology and Paleoanthropology, CAS; Mr. Pan Yaohua and Mr. Wu Junqi, College of Intelligence and Computing, Tianjin University, for

constructive suggestions and help.

Financial support. This research has been supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grants XDA19050101 and XDB26000000) and National Natural Science Foundation of China (61802278).

References

Chen, X., Chen, Q., Zhen, Y., Wang, H., Zhang, L., Zhang, J. and Xiao, Z.: Circumjacent distribution pattern of the Lungmachian graptolitic black shale (early Silurian) on the Yichang Uplift and its peripheral region. Science China Earth Sciences, 61, 1195–1203, 2018.

Chen, X., Zhang, Y., Li, Y., Fan, J., Tang, P., Chen, Q. and Zhang, Y.: Biostratigraphic correlation of the Ordovician black shales in Tarim Basin and its peripheral regions. Science China Earth Sciences, 55, 1230–1237, 2012.

Goldman, D., Sadler, P.M. and Leslie, S.A.: The Ordovician Period, in Geologic Time Scale 2020. Elsevier. p. 631–694, 2020.

Klemme, H.D. and Ulmishek, G.F.: Effective petroleum source rocks of the world: stratigraphic distribution and controlling depositional factors. AAPG Bulletin, 75, 1809–1851. 1991.

Ma, X.:Palaeontology, biostratigraphy and palaeoecology of the graptolite from the Hulo Formation (Darriwilian – Sandbian) in northwestern Zhejiang Province, East China. A Ph.D dissertation submitted to University of Chinese Academy of Sciences (supervised by Prof. Zhang Y-D). 1-301. 2020.

Maletz, J.: Part V, Second Revision, Chapter 13: The history of graptolite classification. Treatise Online, 88:1–11, 2017.

Niu, Z.-B. and Xu, H.-H.: AI-based graptolite identification improve shale gas exploration. bioRxiv. doi: https://doi.org/10.1101/2022.01.17.476477

Peters, S. E. and McClennen, M.: The Paleobiology Database application programming interface. Paleobiology, 42, 1–7, 2016.

Podhalańska, T.: Graptolites–stratigraphic tool in the exploration of zones prospective for the occurrence of unconventional hydrocarbon deposits. Przegląd Geologiczny, 61, 621–629, 2013.

Rahman, I.A., Adcock, K., Garwood, R.J.: Virtual fossils: a new resource for science communication in paleontology. Evolution: Education and Outreach. 5, 635–641, 2012.

Shute, C.H., Foster, T.S.: Curation in museum collections. In: Jones, T.P., Rowe, N.P., eds, Fossil plants and spores: modern techniques. Geological Society of London. 184–186, 1999.

Xu H.H, Nie T., Guo W., Chen Y-S, Yuan W-W.: Palaeontological fossil specimen metadata standard. Acta Palaeotologica Sinica, in press.

Xu, H.-H., Niu, Z.-B. and Chen, Y.-S.: A status report on a section-based stratigraphic and palaeontological database–the Geobiodiversity Database. Earth System Science Data, 12, 3443–3452, 2020.

Xu, H.-H.: High-resolution images of 1550 Ordovician to Silurian graptolite specimens for global correlation and shale gas exploration. https://doi.org/10.5281/zenodo.5205215. 2022.

Zhang, Y.D. Zhan, R.B., Wang, Z.H., Yuan, W., Fang., Liang, Y.,Yan, Wang, Y., Liang, K. et al.: 2020. Illustrations of index fossils from the Ordovician strata in China. Zhejiang University Press. 1–575, 2020.

Zou, C.N., Dong, D., Wang, Y., Li, J., Huang., Wang, S., Guan, Q. et al.: Shale gas in China: Characteristics, challenges and prospects (I). Petroleum Exploration and Development. 42, 689–701, 2015.

Zou, C.N., Gong, J., Wang, H.Y. and Shi, Z.: Importance of graptolite evolution and biostratigraphic calibration on shale gas exploration. China Petroleum Exploration. 24, 1–6, 2019.

Figure 1. Global distribution of graptolite shale and shale gas production region. Most graptolite fossils were yielded from these shale sediments and their distribution is based on their occurrence records in global Ordovician and Silurian sediments. All data are from Peters and McClennen (2016) and Xu et al. (2020). The map is from © OpenStreetMap contributors 2021. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Figure 2. Graptolite species of our dataset are significant to biostratigraphy and dating of Ordovician and Silurian sediments. These graptolites also witnessed several macro-evolution events, including the great Ordovician biodiversity event (GOBE), Late Ordovician mass extinction (LOME).

Radiation of several graptolite groups (bold verticle lines) occurs in this geological time. Two global stratotype sections and points (GSSPs), based on graptolite species record, are in southern China (the spike marks in left figure) (data from Goldman et al., 2020). Bio- or indication zones based on graptolite species assist with identifying mining beds for shale gas exploration in southern China. 16 graptolite indicator-zones are used in the shale gas exploration in China (Zou et al., 2015) (right part in the figure).

Figure 3. The process of creating the graptolite specimen image dataset. The graptolite specimens were carefully curated and revised to select the species with biostratigraphy and application significances. Every image was obtained from specimens that were macro-photographed using a single-lens reflex camera and microscope. After professional revision and cleaning, the whole dataset was uploaded to and stored in our cloud server.

Figure 4. Typical images of graptolite specimens in our dataset. Every image was taken from a unique graptolite specimen. Our dataset only selected the photos that well show morphology of every specimen and diagnostic character of each graptolite species that the specimens represent. The scientific species name of every specimen is given on each image.

Figure 5. Geographic distribution (A) and geologic range (B) of graptolite species of our dataset. Each graptolite specimen locality is represented by a pie chart where each colour is encoded as one graptolite family of the Order Graptoloidea. The sector size is proportional to the specimen number for every family. The radius of the pie chart is proportional to the total number of specimens from the same locality. The dashed-lines circle the main areas of shale gas production. The map is from © OpenStreetMap contributors 2021. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Figure 6. Graptolite species selected as global bio-zone (left) and indicator zone (right) for shale gas favourable exploration beds (FEBs) of our dataset. Among our dataset of 113 graptolite species, there are 22 graptolite index species from global correlation from the Middle Ordovician to (470.0 Ma) to

the Wenlock of the Silurian Period (427.4 Ma), and 16 graptolite species as 'gold callipers' to locate FEBs of shale gas in China. Note that some graptolite species are duplicate in the two lists.

Figure 7. FSIDvis (Fossil Specimen Image Dataset Visualizer) system interface. a) Fossil on geographic distribution view, showing fossil specimen location on the map. The lens (a.1) is a tailor-designed specimens' picker that facilitates users to collect interest fossils of a region where the inner ring and outer ring represent the family and genus. When the user chooses a genus, the corresponding detailed species with images will be listed in the fossil list view (a.2), where the detailed information and further high-~~resolusion~~resolution image if the specimens are given. Hit the space bar for locking the selection. b) Geological age scale view, providing the geologic age selection ability; the top one is the chronostratigraphic age scale, and the bottom one is an age slider that facilitates the users to choose a specific age slot interactively. The web exploration tool of graptolite is provided at http://fsidvis.fossil-ontology.com:8089/. The map is from © OpenStreetMap contributors 2021. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Figure 8. t-SNE embeded~~ing~~ visualization of our graptolite specimen images. Individual specimens are denoted and grouped by different colours~~ and grouped in the visualization~~. These groups ~~also taxonomically~~ match different graptolite families (blocks with several small images).