Cielesta, S., Orlecka-Sikora, B., Staszek, M., Urban, P., Olszewska, D., Ruigrok, E., Toon, S., Picozzi, M., Kwiatek, G., Cesca, S., López Comino, J. A., Isherwood, C., Montcoudiol, N., Jarosławski, J. (2019): SHEER "smart" database: technical note. - *Acta Geophysica*, *67*, 1, pp. 291—297.

DOI: http://doi.org/10.1007/s11600-018-0205-3

Acta Geophysica (2019) 67:291–297
https://doi.org/10.1007/s11600-018-0205-3

RESEARCH ARTICLE - SPECIAL ISSUE

SHEER "smart" database: technical note

Szymon Cielesta[1] · Beata Orlecka-Sikora[1] · Monika Staszek[1] · Paweł Urban[1] · Dorota Olszewska[1] · Elmer Ruigrok[2] · Sam Toon[3] · Matteo Picozzi[4,5] · Grzegorz Kwiatek[6] · Simone Cesca[6] · José Angel López Comino[6] · Catherine Isherwood[7] · Nelly Montcoudiol[8] · Janusz Jarosławski[1]

Received: 1 June 2018 / Accepted: 24 September 2018 / Published online: 28 September 2018

Abstract

The SHEER database brings together a large amount of data of various types: interdisciplinary site data from seven independent episodes, research data and those for the project results dissemination process. This concerns mainly shale gas exploitation test sites, processing procedures, results of data interpretation and recommendations. The smart SHEER database harmonizes data from different fields (geophysical, geochemical, geological, technological, etc.), creates and provides access to an advanced database of case studies of environmental impact indicators associated with shale gas exploitation and exploration, which previously did not exist. A unique component of the SHEER database comes from the monitoring activity performed during the project in one active shale gas exploration and exploitation site at Wysin, Poland, which started from the pre-operational phase. The SHEER database is capable of the adoption of new data such as results of other Work Packages and has developed an over-arching structure for higher-level integration.

Keywords SHEER · Smart database · Multidisciplinary data · Integration · TCS platform

Introduction: the sheer database

The SHale gas Exploration and Exploitation induced Risks (SHEER) smart database gathers a large amount of interdisciplinary data collected from seven independent episodes (Fig. 1), research data and data for the project dissemination process (full list of SHEER publications is available at https://tcs.ah-epos.eu/eprints/view/divisions/SHEER.html). This concerns mainly shale gas exploitation test sites, processing procedures, results of data interpretation and recommendations as well as other documents describing the state of the art. So far, such elements usually have been in the possession of industrial stakeholders (often deemed to be proprietary or sensitive).

The SHEER database harmonizes data from different fields (geophysical, geochemical, geological, technological, etc.) and innovatively creates and provides access to a unique and advanced database of case studies of environmental impact indicators associated with shale gas exploration and exploitation. The data sources within SHEER project are categorized as:

- data sets from shale gas operation sites in Lubocino (Poland) and Preese Hall (UK);
- conventional oil and gas production sites (Groningen, Netherlands and Oklahoma, US); and.
- sites where stimulation for geothermal energy production and geothermal experiments took place.

The reasons for the inclusion of datasets in addition to shale gas developments within the SHEER database were twofold: They have close analogies with the mechanisms

✉ Szymon Cielesta
cielesta@igf.edu.pl

[1] Institute of Geophysics Polish Academy of Sciences, Warsaw, Poland

[2] KNMI - Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

[3] Keele University, Keele, UK

[4] Università degli Studi di Napoli Federico II, Naples, Italy

[5] AMRA S.c. a r.l., Naples, Italy

[6] Helmholtz-Zentrum Potsdam Deutsches GeoForschungs Zentrum GFZ, Potsdam, Germany

[7] RSKW Ltd., Stirling, UK

[8] University of Glasgow, Glasgow, UK

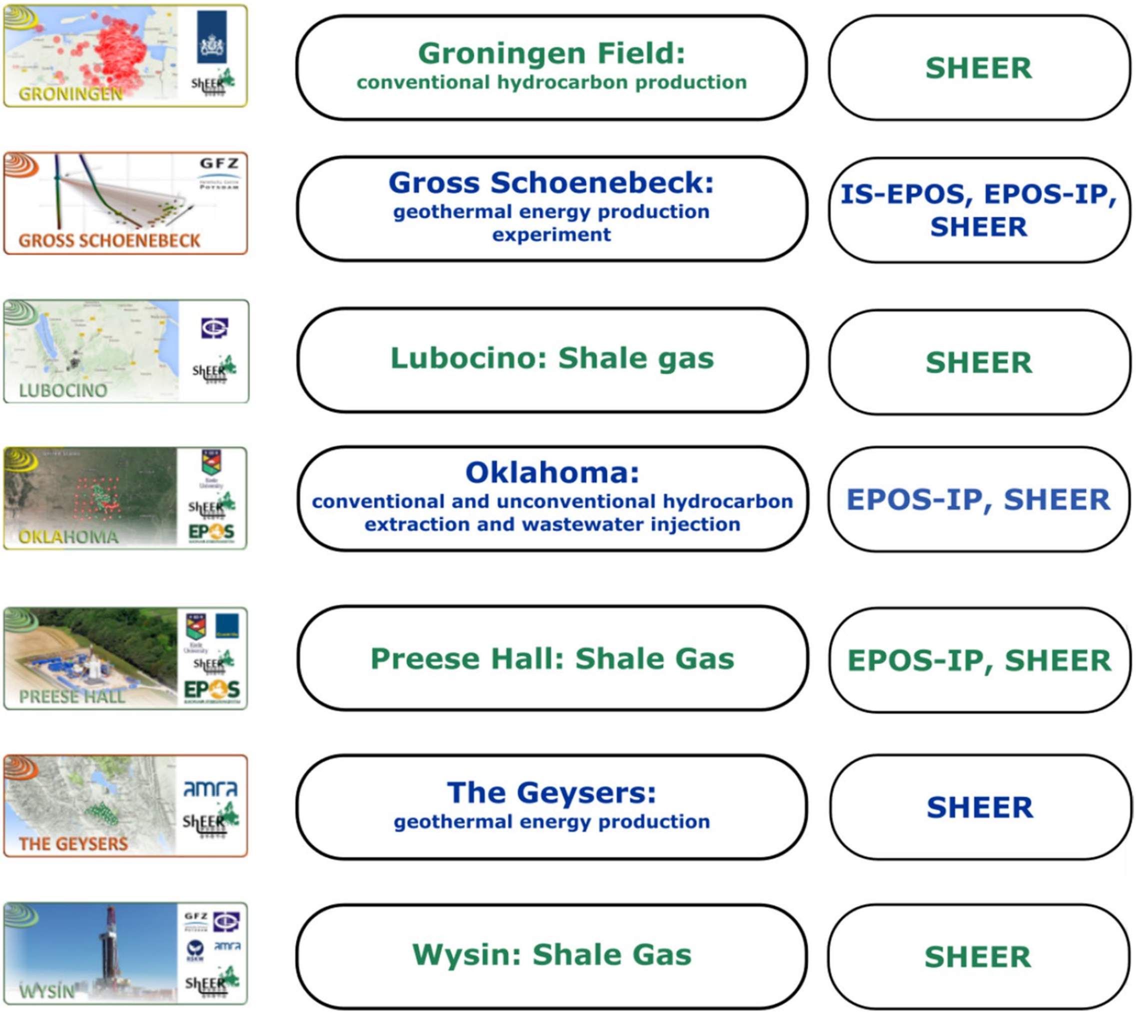

Fig. 1 List of the SHEER database episodes published in IS-EPOS e-platform. (Modified after Cielesta et al. 2018)

involved in shale gas stimulation and induced seismicity problems, and they can be used as proxy.

Besides the "past case studies," an innovative component of the SHEER database is the data from the monitoring activities performed (i.e., in cooperation with Work Package 3—On-site Monitoring) during the SHEER project at an active shale gas exploitation site at Wysin in Pomerania (López-Comino et al. 2017, 2018), where the collection of data and site information started during the pre-operational phase, and therefore, it represents for the scientific community a unique reference database. The SHEER database was therefore designed to be versatile and easily adaptable to new data.

The SHEER database is also a source of lessons learnt during the project: firstly, as a repository of project results dissemination containing work packages reports, deliverables, scientific papers or final guidelines and recommendations, being considered as a project knowledge database, and secondly, as a test bed of IT and database architecture solutions developed already within IS-EPOS project (Digital Research Space of Induced Seismicity for EPOS Purposes, where EPOS stands for European Plate Observing System, http://is-epos.eu/). In turn, presently the SHEER and IS-EPOS databases became the foundations for the next projects databases.

On the way to format-standardized database

Being aware of the multidisciplinary nature of the data going to be integrated within SHEER project, the cooperation with EPOS project was established. The cooperation mainly concerned the Thematic Core Services Anthropogenic Hazards (TCS AH) and the eNode, which integrates research data.

TCS AH aims to integrate distributed research infrastructures (RI) to facilitate and stimulate research on anthropogenic hazards especially those associated with the exploration and exploitation of georesources. These are:

comprehensive data descriptions of case studies, software services and relevant written material.

We started the workflow of data integration by collecting the detailed information on the data types in the relevant questionnaires sent to all data providers gathered in SHEER project. Then, knowing what data we will face, we could start preparing integration plan and, above all, the structures and data types in which they will be stored and made available to SHEER consortium and finally, public. Here, the technologies and standards developed earlier in the IS-EPOS project have been helpful and even proved indispensable.

The process of data preparation, combining harmonization and homogenization and, in the end, integration was additionally monitored by the quality control (QC) workflow. The comprehensive Web services allow speeding up the data integration process and reducing/minimizing the possible mistakes while incorporating diverse heterogeneous datasets.

Database management

Management of such extensive and multidisciplinary observational datasets is much more efficient when the datasets are stored in a database which is easily accessible by the project partners. The SHEER database is stored in the eNODE at the Institute of Geophysics PAS (IG PAS), which is responsible for the technical aspects and for database management. eNODE IG PAS has been created by the CIBIS (Induced Seismicity Research Infrastructure Center) software—the system created to store, manage, configure, verify and describe data in IS-EPOS project (POIG.02.03.00-14-090/13-00).

Presently, there are a two eNodes in TCS AH, one in Poland (CIBIS) and one in France (GeOrchestra) and one Data Center in the Netherlands. During the project, only consortium members had unrestricted access to the SHEER database and all data products, which were worked on via the SHEERWER server (raw data) and IS-EPOS Platform (raw, but also processed data, Kocot et al. 2014; Lasocki et al. 2017). Now, after the project completion, the database is made available according to the rules of the e-Platform, which assumes there is open access from both within and outside the research community. The database stored on SHEERWER is also easily accessible: It could be browsed and downloaded via the sFTP connection or using a webpage with file manager installed. Institutional access is granted to all members of SHEER Consortium.

On eNode, raw data delivered by data providers are stored in separate directories named: “buffer + name of the institution providing data.” Each data provider has an access to buffer directory located on eNode, as well as, to the assigned “QC group.” For QC purposes Redmine, a flexible, open-source project management tool, was used. When the new dataset is provided, the administrator sets a new task in the Redmine system, distributes tasks within Episode Adaptation Centre (EAC) staff and supervises them. Staff verifies, homogenizes and converts (if needed) the raw data from data providers. Data conversion can be performed in the CIBIS system using built-in converters (e.g., seismic).

CIBIS has also separate modules to manage the SHEER database:

- Episodes (which allows to group directories into episodes disregarding to which storage they are assigned);
- Storages (which allows to place directories physically in the server);
- Directories, with internal structure created on the basis of a data scheme (which allows not only adding files and directories but also copying, downloading, compressing, unpacking, renaming, moving or viewing file contents. All versions of uploaded files are being stored in the system, and it is possible to manage these versions.);
- Schemes (which is used to create structure trees inside directories and to define metadata rules and values);
- Data for publication (which is used to set metadata values for selected directories basing on various schema and rules files, which are defined in “Schemes” module).
- Users (which allows to set privileges to the groups of users);
- Configurations (which is used to manage validators and converters. Raw data can be easily converted to chosen format using created converters, e.g., ASCII to mseed converter. Format of integrated data can be easily validated using chosen validators. Smart CIBIS feature is a possibility to add new converters or validators to the already implemented to the system.);

After passing the completeness and quality check, data are placed in final directories according to episode data structure. Each directory (episode) has the same internal structure of subdirectories:

- Data relevant for the considered hazards:
    - Seismic data (e.g., catalog, signals, seismic/ground motion network);
    - Water quality data (e.g., physicochemical water properties, piezometric levels and abstraction rates);
    - Air quality data (e.g., air properties, air stations);
    - Satellite data;
- **Industrial data** (e.g., drilling data, fracture data);
- **Geodata** (e.g., velocity models).

Finally, episodes can be published to IS-EPOS platform if needed.

All acquired data (seismic, operational and environmental) were standardized according to the policies of the EPOS Working Group 10 "Infrastructures for Georesources" (Lasocki et al. 2014), now "Thematic Core Service Anthropogenic Hazards" (tcs.ah-epos.eu), applied on the IS-EPOS platform. In this way, the consistency of all data integrated within IS-EPOS e-Platform and other undertakings within the environmental community are preserved, and their compatibility with already implemented information technology (IT) solutions is ensured.

eNODE/CIBIS cooperates strictly with IS-EPOS e-platform of Thematic Core Service-Anthropogenic Hazards (TCS AH) of EPOS. Mutual communication is based on the LDAP and WebDAV protocol. First one serves for searching for a data/files using metadata, while the second one is used to download the data from LDC. An additional feature of SHEER database regarding metadata creation is that it can be created automatically using patterns supported in CIBIS; there is no need to describe with metadata every file one after the other. Various metadata fields are required depending on the object type (episode, directory or file). Values of metadata are inherited down through the structure of data in the episode, with the possibility to change its value if needed.

Metadata-based approached applied among others in SHEER project was created in agreement with EPOS project. Metadata in EPOS and ESFRI (European Strategic Framework for Research Infrastructures) describe not only datasets, but also users, software and Integrated Core Services (ICS) resources, thus providing a virtualized e-infrastructure environment for geoscientists (Jeffery and Bailo 2014). EPOS TCS data, data products, services and software (DDSS) are provided to the ICS via a communication layer called the interoperability layer. At this level metadata describing all DDSS need to be harmonized into a single metadata catalog which is based on international standards. In EPOS, Common European Research Infrastructure Format (CERIF) is used as an European metadata catalog standard (EPOS—IP WP6 and WP7 Teams 2015).

Multidisciplinary data

The datasets collected in the SHEER database comprehensively describe seismicity, water and air conditions (i.e., data important for hazard assessment), which have been recorded during technological operations and can be used to compare the properties of the aforementioned phenomena with their background states. The technological data are complemented with relevant geodata (Fig. 2).

Due to the multidisciplinary nature of the studies undertaken in the SHEER project, the data collected were heterogeneous. Therefore, the data coming from different research fields (geophysical, geochemical, geological, technological, etc.) had to be homogenized and harmonized. Additionally, the data gathered from past case studies did not conform to a single format. Even the data received from different data providers, although being the same data type, often were prepared in different manner what could be the result of different equipment used for data gathering or just different data processing workflow. This refers to seismic catalog delivered once in.csv, other time in.xls or.txt; having just one type of magnitude, another time having 4 of them. Standardization was applied also to seismic signals delivered alternately as SEED, miniSEED, sometimes just ASCII format.

The remedy to the inhomogeneity of the data types was collecting and storing the data in commonly used standards and formats, e.g., dataless SEED–miniSEED for seismic data, geoTIFF–TIFF for raster geodata and .MAT files for catalogs or technological data.

We highlight that the Generic Data Format (GDF) and Multi-Dimensional Data Format (MDDF) files with standard formats were elaborated as a smart structures to store within the SHEER project measurements results concerning water quality, air quality or other industrial- and geo- data. These data formats are very widely used and easy to use for further processing. GDF and MMDF files contain the identity of the geographic coordinate system in which data are stored, the time zone in which the time is determined, and information about the stored data, such as: unit, data type, names of variables with descriptions.

Within SHEER project the multidisciplinary monitoring in Wysin site has been implemented to jointly assess the short- and long-term risk connected to the most relevant potential hazards of hydrofracturing operations. The environmental monitoring was being led at relatively large area what required setting up a monitoring network. Already mentioned MDDF files were created to meet the requirements of measurements of multi-parameters at different stations (locations). These are expanded GDF files with a more nested structure. The proposed formats are based on the same 9- component, basic data structures that can be easily saved to a file and are easy to manipulate. As GDF examples, air quality or injection rate results can be listed, while, e.g., chemical water properties or water laboratory analyses are being held in MDDF format. Data prepared in this format can be easily converted to ASCII (CSV) format with homogenous structure. Detailed descriptions of the formats used can be found in Project Deliverables available on the Project Web page (www.sheerproject.eu).

Waveforms, which are stored in SeisComp system (SeisComp is providing continuous seismic waveforms in daily files) on SHEERWER, are available via Arclink protocol for

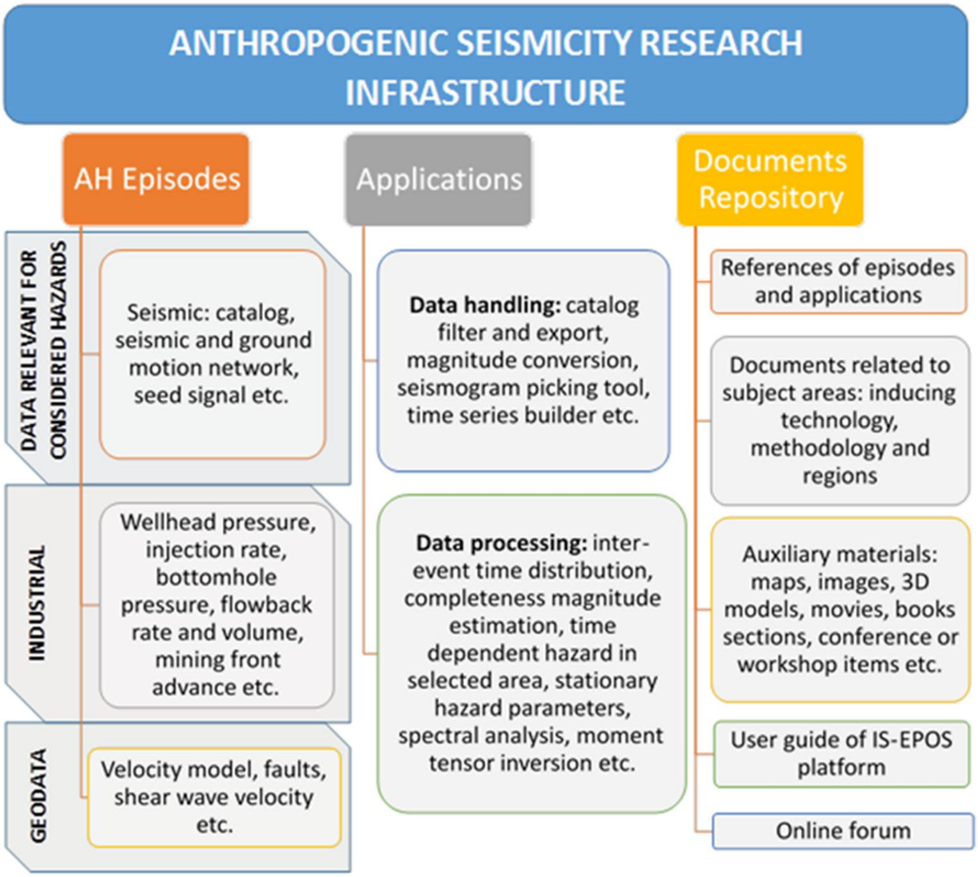
ANTHROPOGENIC SEISMICITY RESEARCH INFRASTRUCTURE
AH Episodes
Applications
Documents Repository
DATA RELEVANT FOR CONSIDERED HAZARDS
Seismic: catalog, seismic and ground motion network, seed signal etc.
INDUSTRIAL
Wellhead pressure, injection rate, bottomhole pressure, flowback rate and volume, mining front advance etc.
GEODATA
Velocity model, faults, shear wave velocity etc.
Data handling: catalog filter and export, magnitude conversion, seismogram picking tool, time series builder etc.
Data processing: inter-event time distribution, completeness magnitude estimation, time dependent hazard in selected area, stationary hazard parameters, spectral analysis, moment tensor inversion etc.
References of episodes and applications
Documents related to subject areas: inducing technology, methodology and regions
Auxiliary materials: maps, images, 3D models, movies, books sections, conference or workshop items etc.
User guide of IS-EPOS platform
Online forum

Fig. 2 Scheme of research infrastructures integrated within IS-EPOS

registered users only. Using Arclink it is possible to download seismic data from any predefined time period and station group in miniSEED or SEED format.

Additionally, the database integrates numerous items of different types (more than 700) with ~500 of them being peer reviewed studies published in international scientific journals. Among the other types of materials stored in the repository, there are books and book chapters, peer reviewed and unreviewed reports, as well as various material from conferences and workshops (e.g., posters and presentations), theses, videos, maps, images, etc. (Rudziński et al. 2017).

Further applications of sheer products

The SHEER project consortium members faced a very difficult challenge, partly as a result of the multidisciplinary aspect of the project, and in consequence, the heterogeneity of the collected data. No less challenging was the integration of data coming from an ongoing Wysin episode with all the results being stored and published almost immediately, in “real time.” This goal was achieved by all of the above-described means, making it possible to call the SHEER database as a smart one.

To preserve and curate the data for long-term use, after the project completion, most of the SHEER database was transferred to the IS-EPOS e-platform, by which means the data are made open access and available to the research community. Access to and sharing of the collected data were recognized by Project Consortium as essential for advancement of science.

The beneficial result of the SHEER database for science is that the collected data and results, which are now disseminated through the IS-EPOS platform, will serve the research community as an open background for the next innovative projects on European shale gas related hazards and risks problem. Accordingly, it is expected that research in this field will gain a broader perspective. In addition, open access to the multidisciplinary data gathered in the framework of SHEER will help to reduce expenditures for further research in this field.

In the already successfully completed projects such as IS-EPOS and SHEER, methodology and tools have been developed to enable the efficient harmonizing, standardizing and integrating process of the data management and final publication. All the developed elements, such as SHEERWER, CIBIS/eNODE and IS-EPOS e-Platform, work together providing smooth communication between themselves and creating a well-designed, smart environment to integrate, manage, process and analyze data for scientific purposes.

The IS-EPOS platform was created as a Science Gateway—a Web-based platform offering to AS researchers a portal for accessing data related to seismicity episodes and performing advanced calculations on these data. The latter was also successfully used within the SHEER project, where, as stated before, all the planned episodes have already been integrated and published.

These tools have already found wide acceptance and appreciation within the scientific community, as confirmed by the over 700 users from 23 countries and over 100 institutions worldwide which have registered on the IS-EPOS e-Platform.

The platform ensures also the opportunity for cooperative work for the SHEER consortium with external projects (Lasocki et al. 2017; Rudziński et al. 2017). Indeed, the external consortium can deposit their data from different projects in a selected data node and get access to these data as well as to all platform facilities through the IS-EPOS portal. Thanks to this facility, the consortium will be able to process the data by means of applications implemented on the IS-EPOS platform and make use of all platform facilities suited for collaboration. Since project data are often embargoed until the end of the project, the accessibility of the data from external projects via the platform may be restricted to the project consortium members only. This matter is addressed by specific agreements between the platform and external project managements. IS-EPOS has been designed in a way which supports and promotes its further development by the means of both AH episodes and applications. Even when not fully accessible, the episodes of a project are visible on the platform for all users, as well as the basic information about the project itself. This approach favors increasing the visibility and outreach of a project. This strategy has already been put in place regarding EPOS IP and SHEER, and will be established soon for the Science for Clean Energy (S4CE) project, and Seismology and Earthquake Engineering Research Infrastructure Alliance for Europe (SERA) project, all funded from the European Union's Horizon 2020 research and innovation program. Over 16 additional AH episodes and more than 10 new high-level application packs are planned to be integrated within the EPOS IP project before September 2019. Within the SERA and S4CE projects, the IS-EPOS platform will serve as a virtual access tool to the integrated AH data and it will be strongly developed in terms of IT solutions facilitating platform usage. Under the framework of some of these projects, non-seismic AH data (e.g., air quality, groundwater chemistry) are in the process of being integrated as well. Within the SHEER project seven AH episodes connected with hydraulic fracturing, geothermal energy production and gas exploitation activities have already been integrated on the IS-EPOS platform and are now available for all registered platform users. Therefore, data quality control, maintenance and safety will be fully managed by TCS. The rules of TCS maintenance and operational costs funding will be prepared through the cooperation with EPOS-ERIC (European Research Infrastructure Consortium).

The IS-EPOS platform sustains an open access tool for the experimentation and training of students and young scientists, who want to perform AS research. Several successful training sessions for undergraduate as well as postgraduate students on AS and hazard have been carried out in Germany, Poland, Sweden and the USA. It is important to emphasize that due to virtual access and eNODE policy, the IS-EPOS platform can be easily used to share and promote the results of any scientific project on AH, where the SHEER project forms the first one, and thereby enrich the scientific community's knowledge and research possibilities. All these functionalities were developed by the collaboration between the EPOS IP and SHEER consortia, which provided a unique partnership within which to test the identified solutions. The Web platform thus promotes new opportunities to study and understand the dynamic and complex solid-Earth system response to human activities by integrating use of data, data products, analysis models, and on-line facilities. IS-EPOS has already been used to facilitate scientific research in recently published papers connected with anthropogenic as well as natural seismicity (e.g., Lizurek 2017; Leptokaropoulos et al. 2017; Gkarlaouni et al. 2017; Rudziński and Dineva 2017; full list of publications acknowledging the IS-EPOS e-Platform is available at http://tcs.ah-epos.eu/eprints/id/saved_search/4).

Acknowledgements This work was supported within SHEER: "Shale Gas Exploration and Exploitation Induced Risks" project funded from Horizon 2020—R&I Framework Programme, call H2020-LCE-2014-1 (www.sheerproject.eu—Grant Agreement No. 640896). The work was also partially supported within statutory activities No 3841/E-41/S/2018 of the Ministry of Science and Higher Education of Poland.

References

Cielesta S, Lasocki S, Leptokaropoulos K, Cesca S (2018) The SHEER approach to shale gas exploration and exploitation associated risks. In: Conference: 7th EAGE workshop on passive seismic, 26–29 March, Krakow, Poland

EPOS – IP WP6 and WP7 Teams (2015) ICS-TCS integration: handbook for tCS integration: level-2, EPOS-IP (Horizon2020—Project number: 676564)

Gkarlaouni C, Lasocki S, Papadimitriou E, Tsaklidis G (2017) Hurst analysis of seismicity in Corinth rift and Mygdonia graben (Greece). Chaos Soliton Fract 96:30–42

Jeffery KG, Bailo D (2014) EPOS: using metadata in geoscience. In: Closs S, Studer R, Garoufallou E, Sicilia MA (eds) Metadata and semantics research. MTSR 2014. Communications in computer and information science, vol 478. Springer, Cham

Kocot J, Szepieniec T, Wójcik P, Trzeciak M, Golik M, Grabarczyk T, Siejkowski H, Sterzel M (2014) A framework for domain-specific science gateways. E-science Distrib Comput Infrastruct 8500:130–146. https://doi.org/10.1007/978-3-319-10894-0_10

Lasocki S et al (2014) White paper of WG10 "INFRASTRUCTURE FOR GEORESOURCES" Thematic Core Service Anthropogenic Hazard—interactive research infrastructure for solid-earth by S. Lasocki, B. Orlecka-Sikora, J.R. Grasso and the members of EPOS WG10 "Infrastructures For Georesources". 20 Sep 2014

Lasocki S, Orlecka Sikora B, Leptokaropoulos K, Sterzel M, Szepieniec T, Kocot J, Mutke G, Barański A, the IS-EPOS team (2017) IS-EPOS: a digital research space to facilitate integrated approach to anthropogenic seismic hazards. In: Proceedings of the 16th world conference on earthquake engineering, 16WCEE, Santiago, Chile

Leptokaropoulos K, Staszek M, Cielesta S, Urban P, Olszewska D, Lizurek G (2017) Time dependent seismic hazard in Bobrek coal mine, Poland, assuming different magnitude distribution estimations. Acta Geophys. https://doi.org/10.1007/s11600-016-0002-9

Lizurek G (2017) Full moment tensor inversion as a practical tool in case of discrimination of tectonic and anthropogenic seismicity in Poland. Pure Appl Geophys 174:197–212

López-Comino JA, Cesca S, Kriegerowski M, Heimann S, Dahm T, Mirek J, Lasocki S (2017) Monitoring performance using synthetic data for induced microseismicity by hydrofracking at the Wysin site (Poland). Geophys J Int 210:42–55

López-Comino JA, Cesca S, Jarosławski J, Montcoudiol N, Heimann S, Dahm T, Lasocki S, Gunning A, Capuano P, Ellsworth WL (2018) Induced seismicity response of hydraulic fracturing: results of a multidisciplinary monitoring at the Wysin site, Poland. Sci Rep. https://doi.org/10.1038/s41598-018-26970-9

Rudziński Ł, Dineva S (2017) Towards energy magnitude for mining induced seismicity. In: Proceedings of the 9th international symposium on rockbursts and seismicity in Mines, RaSiM9, Santiago, Chile

Rudziński Ł, Lasocki S, Orlecka-Sikora B, Szepieniec T, Leptokaropoulos K, Garcia-Aristizabal A, Grasso J-R, Kwiatek G, IS-EPOS and EPOS-IP teams (2017) IS-EPOS platform for anthropogenic seismicity research: a modern e-tool of EPOS Thematic Core Service "Anthropogenic Hazards", designed to integrate infrastructures and facilitate research of human induced seismicity. In: Vallejos JA (ed) Proceedings of the 9th international symposium on rockbursts and seismicity in mines—RaSiM9, Nov 15–17, Santiago, Chile. Editec S.A., Santiago, Chile, pp 15–22